NPS55-81-017

NAVAL POSTGRADUATE SCHOOL

Monterey, California

EFFECT OF TASK DURATION ON VOICE RECOGNITION SYSTEM PERFORMANCE

by

John W. Armstrong
and
Gary K. Poock

September 1981

Prepared for:

Naval Electronic Systems Command
Washington, D.C. 20360

NAVAL POSTGRADUATE SCHOOL
MONTEREY, CALIFORNIA

Rear Admiral J. J. Ekelund
Superintendent

D. A. Schrady
Acting Provost

This investigation was sponsored by Mr. Frank Deckelman, NAVELEX, Code 330. The work was performed by the authors at the Naval Postgraduate School, Monterey, California.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: NPS55-81-017
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (*and Subtitle*): Effect of Task Duration on Voice Recognition System Performance
5. TYPE OF REPORT & PERIOD COVERED: Technical
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): J. W. Armstrong and Gary K. Poock
8. CONTRACT OR GRANT NUMBER(s):
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Naval Postgraduate School, Monterey, CA 93940
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: N0003980WR09041; PE 62721N
11. CONTROLLING OFFICE NAME AND ADDRESS: Naval Postgraduate School, Monterey, CA 93940
12. REPORT DATE: September 1981
13. NUMBER OF PAGES: 71
14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*): Naval Electronic Systems Command, Washington, D.C. 20360
15. SECURITY CLASS. (*of this report*): Unclassified
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:
16. DISTRIBUTION STATEMENT (*of this Report*): Approved for public release; distribution unlimited.
17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*): VTAG, Voice Recognition, Automatic Word Recognition, Mental Loading, Effects of Task Duration
20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*):

This report describes an experiment examining voice recognition performance when subjects were required to make a decision every 1.5 seconds for 20 consecutive minutes while speaking to a voice recognizer at the same time. Results showed mental loading degraded voice recognition performance. Results also showed performance degradation similar to a vigilance decrement over eight trials of 2.5 minutes each.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

FOREWORD

This investigation was sponsored by Mr. Frank Deckelman, NAVELEX, Code 330. The work was performed by the authors at the Naval Postgraduate School, Monterey, California.

This report is one of a series concerned with the possible applications of using voice recognition technology in command and control tasks. A listing of other reports in this series is contained in Appendix M.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EFFECT OF TASK DURATION ON VOICE RECOGNITION SYSTEM PERFORMANCE

A. INTRODUCTION

This experiment was conducted to substantiate and further investigate a phenomenon observed by Armstrong and Poock (1981) in which voice recognition system performance degraded significantly over a period of five minutes. Specifically, recognition error rates in that study were statistically greater in the second half of a five minute trial than in the first half. Obviously, with only two time periods in that study, it was impossible to determine if performance would have continued to degrade or would have leveled off if the duration of the task had been longer.

B. OBJECTIVE

The objective of this experiment was to determine the effect of task duration on performance of a voice recognition system comprised of a human operator and a discrete utterance voice recognition system.

In contrast to the earlier study by Armstrong and Poock (1981) in which two time periods of 2.5 minutes each were used, this study would use 8 periods of 2.5 minutes each for a total task duration of 20 minutes. In addition, two experimental conditions would be used examining mental loading versus no mental loading over the duration of the voice recognition task.

C. SUBJECTS

Twenty-four subjects participated on a volunteer basis with no monetary or other incentive. Twenty-three of the subjects were students at the Naval Postgraduate School (NPS). They included 20 male military officers representing the United States Navy, Army, Marine Corps and Coast Guard,

one female military officer of the United States Navy, and one male civilian from the United States National Security Agency. One subject was a male civilian staff member at NPS. All subjects were between the age of 26 and 38 inclusive and the ranks of the military officers ranged from Lieutenant to Lieutenant-Commander and from Captain to Major inclusive.

Only two of the subjects had any previous experience on the RATER - two hours and four hours. Only three of the subjects had any previous experience on the voice recognition system used in the experiment - 0.5, 2 and 30 hours.

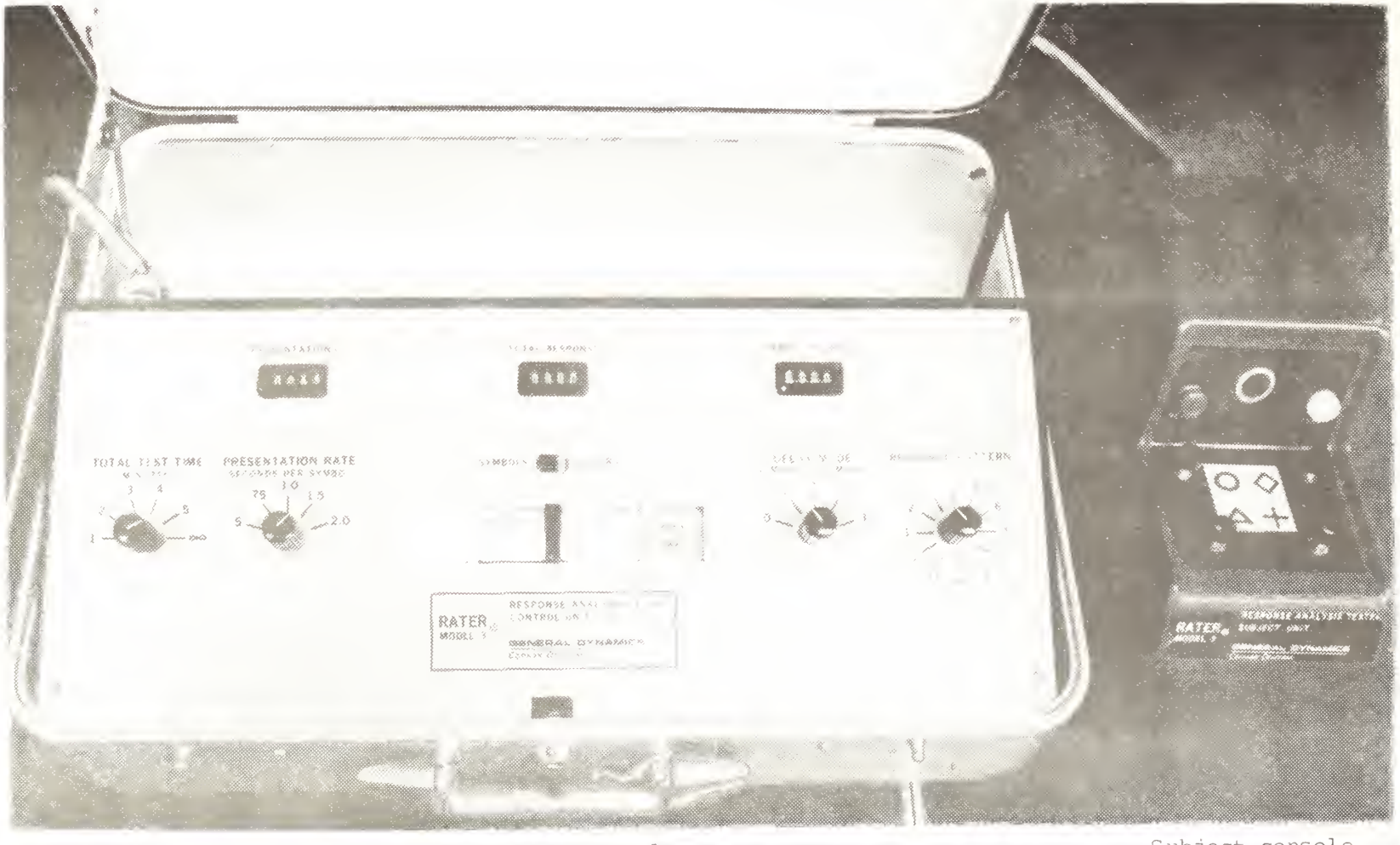
D. EQUIPMENT AND VOCABULARY USED

The same equipment used by Armstrong and Poock (1981) was used in this experiment and is briefly described here.

1. Response Analysis Tester (RATER)

The General Dynamics Response Analysis Tester (RATER, Model 3) shown in figure 1 was used to simulate operator mental loading. Brady (1968) described the RATER as a "psychomotor testing instrument designed to provide sensitive, reliable measurement of any impairment of response speed/accuracy and short-term memory for patterned or color stimuli." Long and Fishburne (1973) provide normative RATER performance data for a student naval aviator population and reference several studies in which the RATER was used. Newsom, Brady and O'Laughlin's study (1966) of performance in a revolving space station simulator found that turning the head while in a rotating environment resulted in degraded short term memory as measured on the RATER.

The RATER consisted of a small subject console which contained a display window and four response buttons in a two by two arrangement and a larger experimenter console which contained the controls and digital counters. These counters were used in the derivation of subject RATER performance data.



Experimenter console

Subject console

FIGURE 1. RESPONSE ANALYSIS TESTER (RATER)

The RATER was used to generate and display random sequences of four individual stimulus symbols - a triangle, a circle, a cross and a diamond - in the window of the subject console. Stimulus symbols were presented at a constant rate of one symbol every 1.5 seconds. Four response buttons on the subject console were associated with each of the four symbols and labelled accordingly.

Based on the results of Armstrong and Poock (1981), the mental loading condition of delay one was used in this experiment.

While the n^{th} stimulus of the sequence, St(n), was being displayed and before St(n+1) replaced it 1.5 seconds later, the subject was required to press the correct response button in order to score a correct response. In delay one the correct response button for the n^{th} stimulus was the one which corresponded to the stimulus symbol comprising St(n-1). In other words, when a stimulus symbol appeared, the correct response was for the symbol which had just previously been displayed. The correct response was never the response button for what was currently being displayed, but rather the correct response was always the response for the stimulus symbol which had appeared "one back".

The RATER was used solely as a device to load the subjects mentally, i.e. to load the subjects through tasking which was primarily decision-making in nature.

2. Voice Recognition System and Choice of Vocabulary

A Threshold Technology Inc. Model T600 discrete utterance voice recognition system (which will hereafter be referred to as the T600) was used as the equipment component of the combined equipment plus human operator voice recognition system. A description of the operation and capabilities of this equipment was provided in Armstrong and Poock (1981), Poock (1980) and Armstrong (1980).

The vocabulary used in this experiment consisted of 50 different utterances. Thirty were single words selected by the experimenter from the Listener's Answer Sheets of the Modified Rhyme Test, one of the four test types which have been commonly used in measuring intelligibility in speech communication (Kryter, 1972). Sixteen of these 30 words were eight pairs of rhyming words which, within each pair, differed only with respect to initial consonant - for example, "beat" and "peat". The other 14 words were seven pairs of non-rhyming but similar words which, within each pair, differed only with respect to final consonant - for example, "sap" and "sat". The other 20 utterances were chosen by the experimenter from single words commonly used in Command and Control environments; they were chosen to be more easily distinguished from each other and from the other 30 words of the vocabulary.

All words of the vocabulary were one or two syllables in length. Short words were deliberately selected to facilitate generation of as many T600 word recognition attempts as possible in the limited time that each volunteer subject was available. The vocabulary is listed by word type in Appendix A. A listing in the order in which the words were trained is attached to the written instructions initially given to subjects and is contained in Appendix C.

This particular vocabulary was chosen to increase the likelihood of recognition errors by the T600 for the following reason. (T600 recognition errors (RE's) are operationally defined in the Dependent Variables section). Recognition accuracy with older Threshold Technology Inc. voice recognition equipment similar to the T600 and using more normal vocabularies (i.e. comprised entirely of more easily distinguished words) has often been better than 99%, as for example, in the studies by Martin and Grunza (1974),

Scott (1975) and Scott (1978). This level of accuracy would produce an average of about one (or less) RE's per 100 spoken utterances. It was anticipated that if operator mental loading did affect recognition accuracy then the effect would be relatively small and, due to the discrete nature of RE's, would probably not be easily distinguishable if only one RE per 100 utterances were being observed - for example, a 20% increase in RE's would probably not be great enough to produce a sufficient number of increased RE observations to be statistically distinguishable from inherent random variation. However, if a vocabulary could be chosen to produce approximately ten RE's per 100 utterances a 20% increase in RE's should be more easily distinguishable as this would result in an average observation of 12 RE's per hundred utterances.

An alternative method of detecting a small expected change in recognition accuracy would be to increase the number of utterances spoken by the subjects. This was not considered feasible here because of the greatly increased time which would be required of each of the volunteer subjects; the experimental design used required between one and two hours per subject. For this reason the former method, special vocabulary, was used.

3. Arrangement of Equipment used

Figure 2 illustrates the functional relationships among the various experimental devices used in the experiment. A photograph of the experimenter control station is shown in figure 3. The subjects were seated one at a time in an Industrial Acoustics Co. Inc. Controlled Acoustic Environments booth. The subject console of the RATER was on a table in front of the subject.

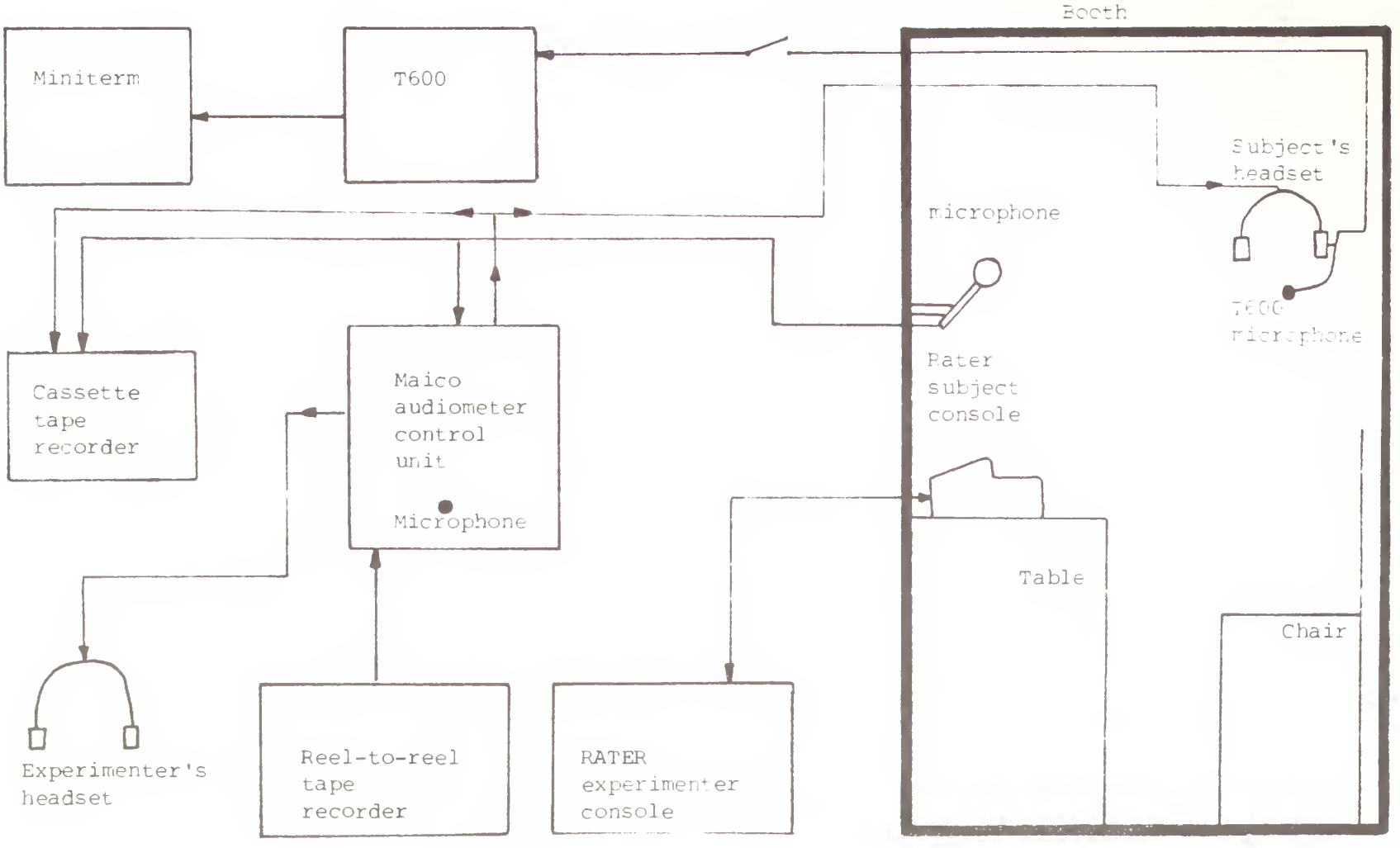


FIGURE 2. BLOCK DIAGRAM OF EXPERIMENTAL CONTROL SYSTEM

Miniterm T600 Cassette tape recorder Maico Audiometer RATER (experimenter console)

FIGURE 3. EXPERIMENTER CONTROL STATION

A Maico Model MA-24B Dual Channel Research and Diagnostic Audiometer and headsets were used to provide oral communication between the subject and the experimenter. The experimenter could speak to the subject by depressing a "talk-over" switch. Another microphone, placed in the booth, was live at all times and permitted the experimenter to hear what was happening in the booth - in particular, what the subject said. A Sony model TC 124 cassette tape recorder was connected to permit simultaneous recording of the signals detected by the booth microphone and the signals that the subject received over his headset.

The special T600 system noise-cancelling microphone was mounted on the subject's headset and connected only to the T600. The microphone ON/OFF switch was located outside of the booth.

A Computer Devices Inc. Model 1203 Miniterm portable terminal was connected to the T600 system in such a manner that when the T600 recognized an utterance the output string for that utterance was typed at the terminal. The T600 was programmed so that the ASCII output stream associated with each utterance of the vocabulary was simply the letters spelling the utterance followed by a carriage return and a line feed; thus, for example, if in the recognition mode the T600 "thought" that a subject said "attack", the word "attack" was displayed on the CRT on a separate line and printed at the terminal, also on a separate line. This provided the experimenter with a paper printout of T600 recognition activity which, with the correct utterances recorded on the cassette tape recorder, permitted thorough analysis of the data. Accurate, manual, real-time analysis by the experimenter using only the T600 CRT was infeasible primarily because of the rate at which the T600 was required to process signals for recognition - one word every three seconds.

An Akai model 4000DS Mk II reel-to-reel tape recorder was connected to the Maico Audiometer and used to present stimuli to the subject.

E. EXPERIMENTAL PROCEDURE

Subjects were tested one at a time during normal working hours. They were first required to complete the Subject Data Form (Appendix B) and then read the pages of written instructions (Appendix C) which briefly introduced the experiment and provided general guidelines on inputting voice data to the T600. Remaining instructions to the subject were given orally by the experimenter.

Subjects were next given a brief demonstration of the operation of the T600. For this stage the T600 microphone and the headset on which it was mounted were removed from the booth and the microphone was reconnected outside of the booth so that the subject could immediately see what happened when speech signals were input to the T600. The importance of the guidelines which the subject had just read were demonstrated during this stage and the subject was allowed to familiarize himself with the T600 for about five minutes.

The T600 microphone and the headset on which it was mounted were then reconnected inside the booth.

The 50 word vocabulary was then trained one word at a time. The experimenter had all of the T600 controls outside of the booth and closely controlled the training process, requiring the subject to retrain words as necessary - for example, if a word was initially trained monotonously. The T600 was next put in the recognition mode and recognition of each word of the vocabulary was checked. Words which initially could not be recognized were retrained until they could be correctly recognized. If

a word was correctly recognized immediately it was not checked further. Words not correctly recognized immediately were retrained if more than one recognition error was obtained in three attempted recognitions of the word. Retrained words were rechecked and retrained again as necessary.

The subject next received, via his headset, a 2.5 minute tape recording of the 50 words of the vocabulary arranged in random order and presented at a constant rate of one word every three seconds. The subject was instructed to repeat the words one at a time for recognition by the T600. He was advised to try to repeat each word and to guess with a word in the vocabulary if he was uncertain.

Next the subject was briefed on the RATER task that he would be performing - delay one. He was advised that his RATER scoring would be number of correct responses minus number of incorrect responses, which included both omission and commission errors. The subject was also advised that he was not required to attain any particular proficiency level on the RATER but that it was sufficient that he understood the task and did his best. He was then allowed to practice the RATER task for up to 10 minutes. The RATER was used in the self-pace mode during parts of the practice if requested by the subject. In the self-pace mode the symbol displayed was replaced by the next symbol in the sequence only when a correct response was made.

When the subject advised the experimenter that he no longer wished to practice on the RATER, the subject was given a combined 2.5 minute RATER delay one plus the word repetition for voice recognition practice. The subject was played the same 2.5 minute tape recording that he had heard earlier and was instructed as before to repeat the words one at a time

for recognition by the T600. He was advised that this was the higher priority task but that he was to simultaneously perform the RATER task as well as he could with whatever capabilities he had remaining after attending to the priority task. The subject was also reminded to be sure to repeat each of the taped words and to guess with a word in the vocabulary if he was uncertain.

The subject was then exposed to the two experimental conditions of this experiment:

1) One experimental condition was only for voice recognition, in which for 20 minutes, the subject heard a vocabulary word said to him over his headset once every 3 seconds. Upon hearing the word, he would say it for voice recognition. This involved no RATER task and will be called condition NRT (no RATER task).
2) The other experimental condition was the same as the previously described NRT condition, but in addition, every 1.5 seconds for 20 minutes a new visual display was being presented by the RATER in delay one so that the correct response was always the stimulus symbol which preceded the one being displayed. This condition was called RD1 (RATER delay 1). The second condition, RD1, therefore required the subject to speak every 3 seconds, but also imposed a mental load where he had to simultaneously make a decision on the RATER task every 1.5 seconds and make a push button response. (In the experiment, eight different random orderings of the 50 word vocabulary were strung together to make one continuous 20 minute presentation of words to the subject's headset.)

Subjects were reminded that the repetition of words for recognition by the T600 was the higher priority task and to guess with a word from the vocabulary if they were uncertain, as during the practice. (The purpose of this instruction was to ensure that the T600 received the same, or at least nearly the same, utterances for recognition during each trial half and thus provide a common basis for comparison of T600 recognition errors.) By monitoring the T600 CRT display and RATER counters, listening to booth activity via the booth microphone, and post-experiment questioning of subjects, the experimenter ensured that subjects adhered to the instructions that they had been given.

Immediately after a subject completed each condition, and before he was allowed to leave the booth, he was instructed to complete the "Feeling Tone Checklist" shown in Appendix D in accordance with the instructions also shown in Appendix D. This checklist, developed by Pearson and Byars (1956), was administered to assess possible differential subjective fatigue after each of the four different mental loading conditions.

During the experimental conditions subjects were not given feedback on their RATER performance. During the practice sessions the only feedback given to subjects regarding T600 recognition of their speech was the knowledge of which words required retraining; no feedback regarding T600 recognition performance was given to subjects during the experimental conditions. Those subjects who indicated interest on their "Subject Data Sheets" were individually briefed immediately after they completed the last experimental condition concerning their RATER performance, T600 recognition of their speech and the hypotheses being tested.

Subjects were allowed to take short rest breaks as they wished during the training and practice sessions and before each of the experimental

conditions. A drinking fountain was located nearby for any subjects who became thirsty or whose throats became dry.

Subjects' watches were removed before the two experimental conditions so that subjects did not know exactly how much time remained. This was done to avoid possible endspurt effects. In the case of the experimental condition corresponding to condition NRT in this experiment, it was felt that as task duration increased the subject would tend to become bored (under-aroused) and consequently performance of the voice recognition system would degrade. It was felt that the literature on human vigilance performance (for example, Davies and Tune, 1969) would give a rough indication of how performance would degrade with task duration.

Although it could be argued that the voice input task of condition NRT was not properly a vigilance task because of the high and constant rate of presentation of signals, in this case 20 words for the subject to repeat each minute, other tasks involving rates of signal presentation comparable to those of the current experiment have been considered vigilance tasks; for example, Kennedy (1972) used up to 19 signals per minute. In any event, consulting the vigilance literature was considered valid because the purpose of this was not to predict precisely the performance degradation but simply to get a rough indication of how such degradation would depend on task duration. The tentative conclusion reached was that system performance initially would degrade quickly but then gradually approach some asymptotic level; how long performance would degrade significantly and how quickly this asymptotic level would be approached could not be estimated in advance.

In the case of the experimental condition corresponding to condition RD1 of the experiment, it was felt that as task duration increased subjects would tend to become fatigued and consequently performance would degrade.

However, as task duration increased subjects might learn to cope better with performing the two tasks simultaneously; this would tend to increase performance with task duration. Overall, it was felt that system performance initially would degrade quickly as observed by Armstrong and Poock (1981) but then eventually approach some asymptotic level; again, how long performance would degrade significantly and how quickly this asymptotic level would be approached could not be estimated in advance.

F. DEPENDENT VARIABLES

The following were calculated for each of the eight 2.5 minute time periods comprising each experimental condition.

1. T600 recognition errors (RE's)
2. Subject verbal errors.

In this experiment a T600 recognition error was operationally defined to be a failure of the T600 to recognize correctly any vocabulary word which a subject said; this included both incorrect recognition (for example, the subject said "beat" and the T600 "thought" he said "peat") and rejection (for example, the subject said "dip" and the T600 failed to recognize it and emitted a "beep" sound). This definition is different from most definitions of recognition error in the voice recognition literature which do not include rejections - for example, Martin and Grunza (1974). The operational definition used in this experiment was considered more consistent with the aim of this research - i.e. to answer the question: Would increased operator mental workload (with respect to that experienced during training of the recognition device) result in changes in his speech which would in turn result in degraded performance of the voice recognition system? It was believed that if the T600 rejected "dip" when said by a subject under condition RD1,

but not when said by the same subject under condition NRT, this suggested changes in system performance as a result of changes in the subject's speech and accordingly should be recorded and analyzed.

A subject verbal error was defined as a failure of the subject to repeat correctly the presented word. This failure could be either a failure to respond (omission) or responding with a non-vocabulary word or the wrong vocabulary word (commission).

G. HYPOTHESES

The following hypotheses were to be tested.

1. Hypotheses Regarding T600 Performance

a. H_o: T600 recognition error rate (RER) would be the same in conditions NRT and RD1.

H_1: H_o false.

It was expected that RER(RD1) would be greater than RER(NRT), as RER(RD1) was greater than RER(NRT) in the experiment of Armstrong and Poock (1981).

b. H_o: T600 recognition error rate would be the same in each of the eight consecutive 2.5 minute time periods comprising each experimental condition.

H_1: H_o false.

It was expected that recognition error rate would tend to increase with task duration, as discussed earlier.

2. Hypotheses Regarding Subject Performance

a. H_o: Subject verbal error rate (VER) would be the same in conditions NRT and RD1.

H_1: H_o false.

It was expected that VER(RD1) would be greater than VER(NRT).

b. H_o: Subject verbal error rate would be the same in each of the eight consecutive 2.5 minute time periods comprising each experimental condition.

H_1: H_o false.

It was expected that subject verbal error rate would tend to increase with task duration, as discussed earlier.

c. H_o: RATER scores would be the same during each of the eight consecutive 2.5 minute time periods comprising experimental condition RD1.

H_1: H_o false.

It was expected that RATER scores would tend to decrease with task duration, as discussed earlier. (RATER performance was recorded at the end of each 2.5 minute time period.)

d. H_o: Subject subjective fatigue (as measured by the "Feeling Tone Checklist" of Pearson and Byars, 1956) would be the same for both experimental conditions, NRT and RD1.

H_1: H_o false.

It was expected that condition RD1 would induce more fatigue than condition NRT.

H. EXPERIMENTAL DESIGN

A conceptual design of the experiment is shown in figure 4. This is a three factor factorial design.

Twelve of the 24 subjects received condition NRT first and condition RD1 last; the other 12 subjects received condition RD1 first and condition NRT last. Subject to this restriction, the order of presentation of the two conditions to any particular subject was assigned randomly.

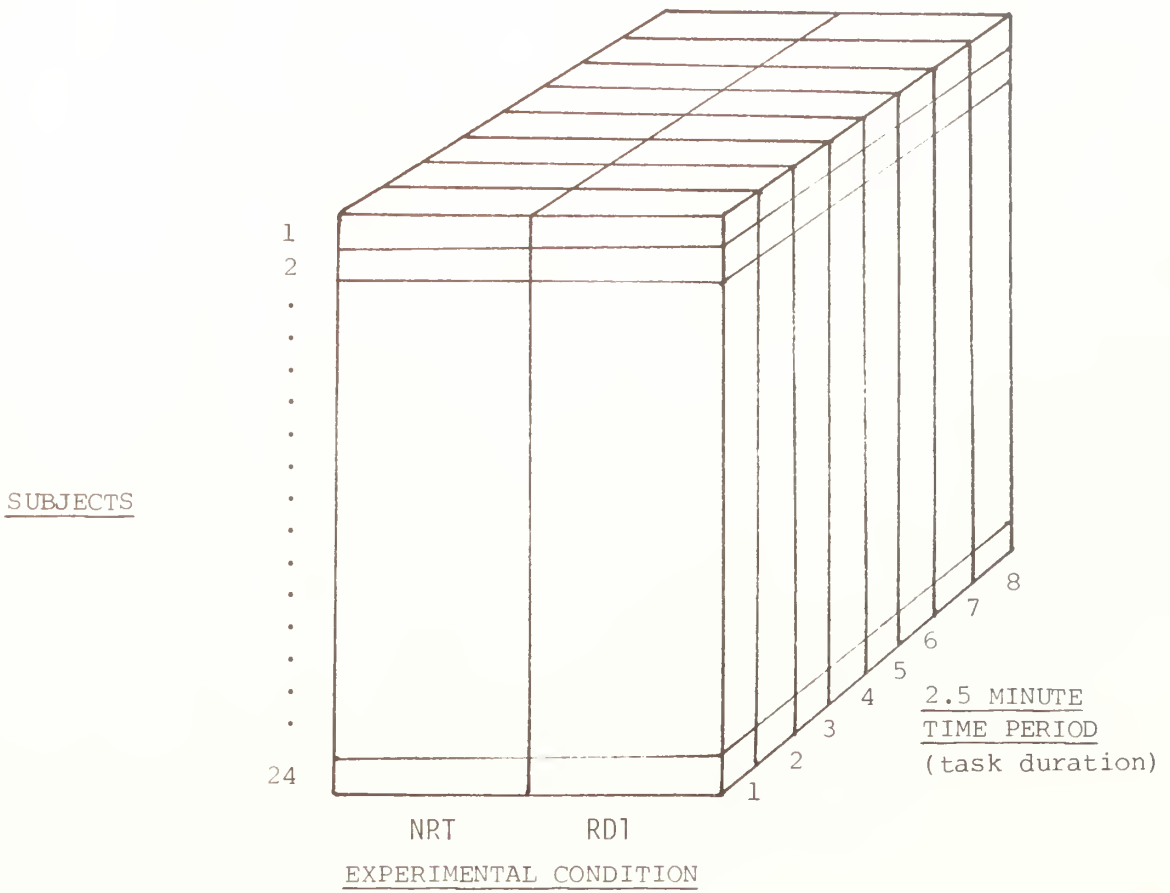


FIGURE 4. CONCEPTUAL DESIGN OF THE EXPERIMENT

An α of .10 was selected for testing the hypotheses as was used by Armstrong and Poock (1981).

I. RESULTS

1. Results for T600 Performance

Appendix E shows total T600 recognition errors for each subject for each 2.5 minute time period of each experimental condition. Mean T600 recognition error rates for each 2.5 minute time period, experimental condition and vocabulary word type, expressed in recognition errors per 100 spoken utterances, are shown in Table I.

Figures 5 and 6 are plots of the recognition error rate observations for conditions NRT and RD1 respectively. Figures 7 and 8 are plots of the arcsin transformed recognition error rate observations for conditions NRT and RD1 respectively. Figures 7 and 8 show that the parametric analysis of variance homogeneity of variance assumption was adequately met. Since the parametric analysis of variance is quite robust regarding its Normality assumption (Scheffé, 1959), it was felt that this assumption also was adequately met and a parametric analysis of variance was performed on the arcsin transformed data. The results are summarized in Table II. The model for this analysis was:

$$Y_{ijk} = u + C_i + T_j + S_k + CT_{ij} + e_{ijk}$$

where Y_{ijk} = arcsin transformed recognition error rate for experimental condition i, 2.5 minute time period j, and subject k; the range of Y_{ijk} is 0 to π.

u = common experimental contribution to Y_{ijk}

C_i = contribution of experimental condition i, i = 1, 2 (NRT, RD1)

TABLE I

MEAN T600 RECOGNITION ERROR RATES*

BY EXPERIMENTAL CONDITION

NRT	13.53%
RD1	15.52%

BY 2.5 MINUTE TIME PERIOD

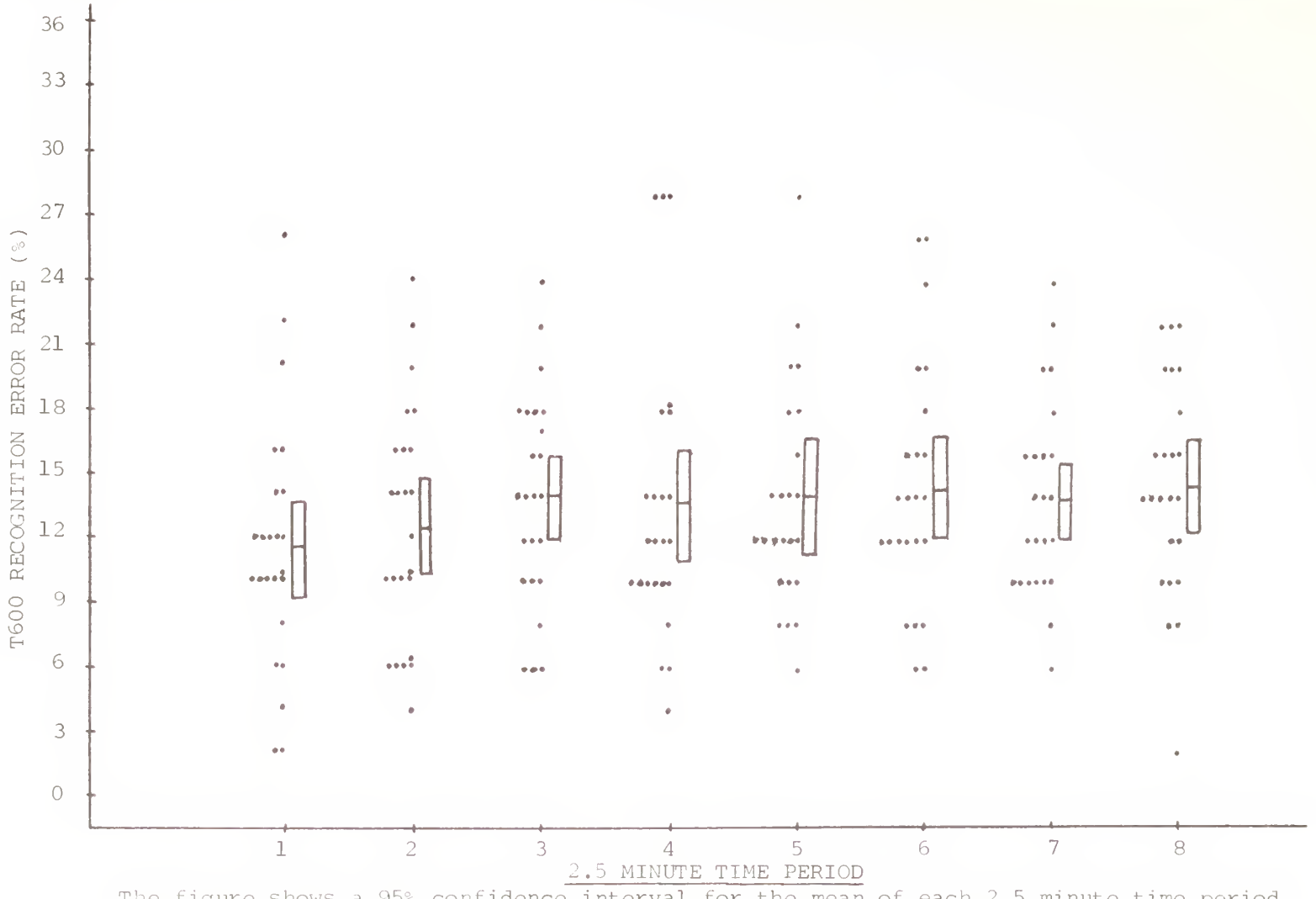
1	12.31%
2	13.93%
3	14.44%
4	14.88%
5	14.55%
6	15.62%
7	14.42%
8	16.03%

BY VOCABULARY WORD TYPE

Rhyming	27.90%
Non-rhyming but similar	14.45%
Operational	3.83%

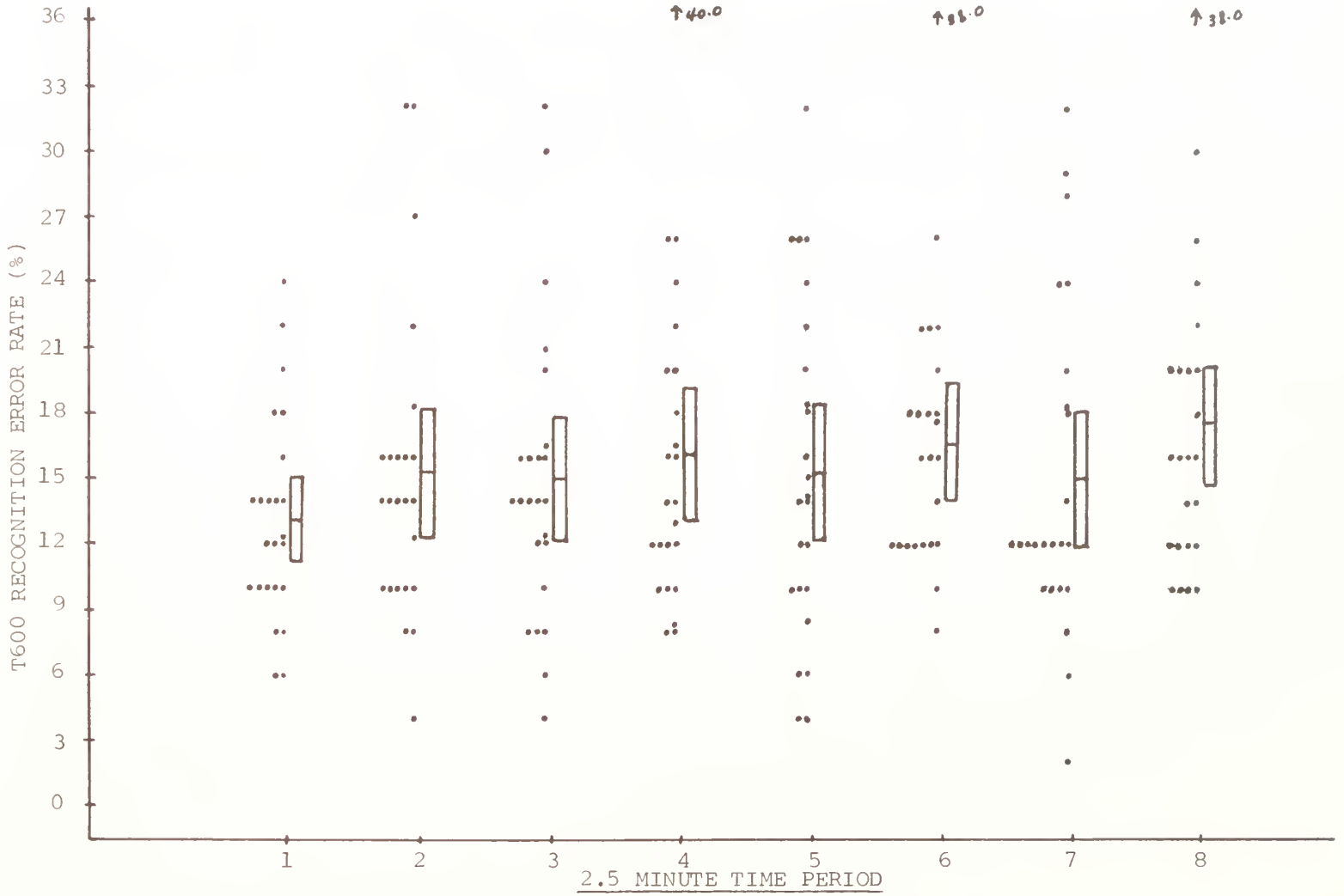
OVERALL 14.52%

* Expressed in recognition errors per 100 spoken utterances. A recognition error was operationally defined in this research to be a failure of the T600 to recognize correctly any vocabulary word which S spoke and includes both incorrect recognition and rejection of vocabulary words; recognition errors do not include those cases where S spoke a word not in the vocabulary (or coughed, sighed, etc.) and the T600 generated a recognition.



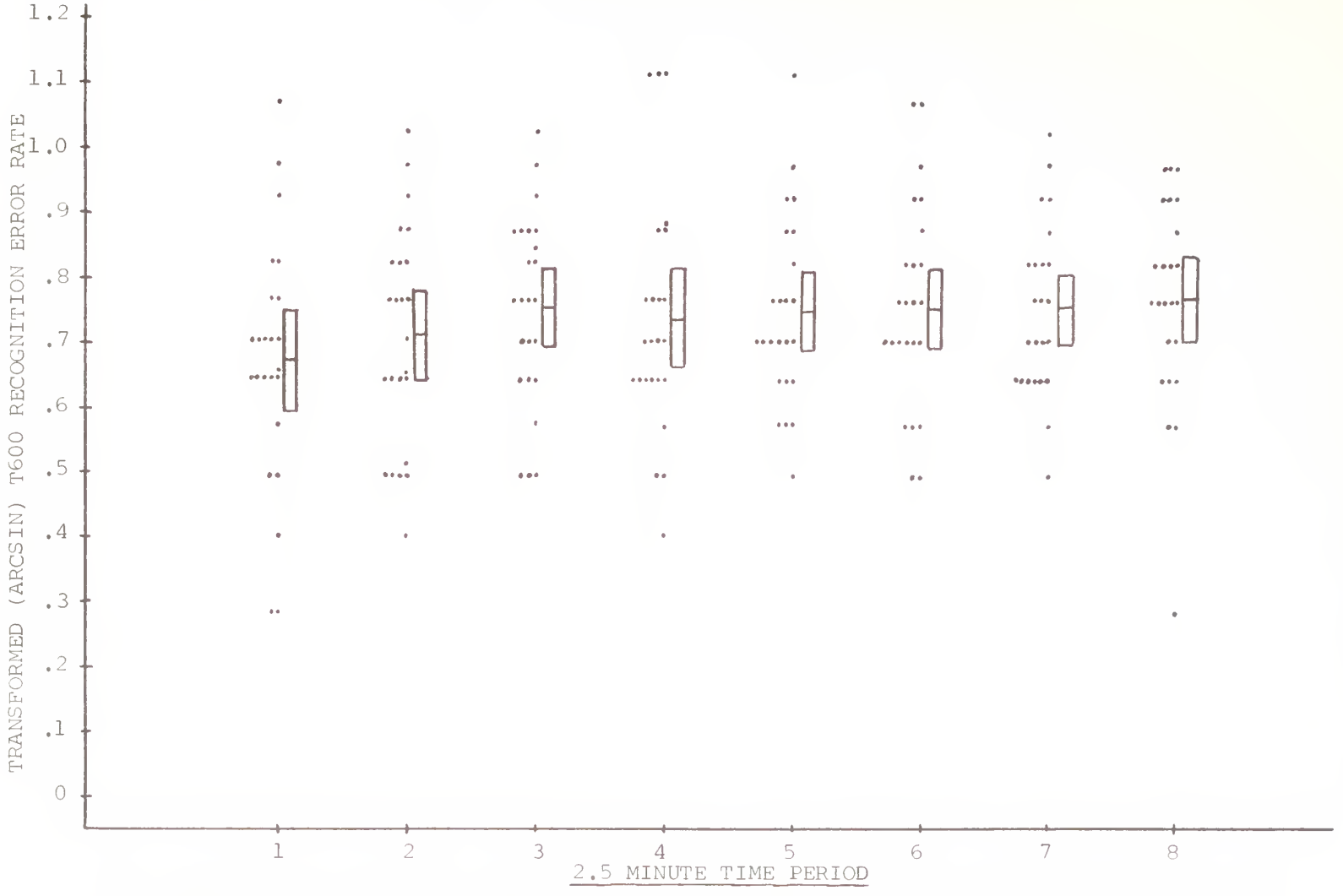
The figure shows a 95% confidence interval for the mean of each 2.5 minute time period. A level of significance, α, of .10 was selected in the experimental design.

FIGURE 5. T600 RECOGNITION ERROR RATE OBSERVATIONS: CONDITION NRT



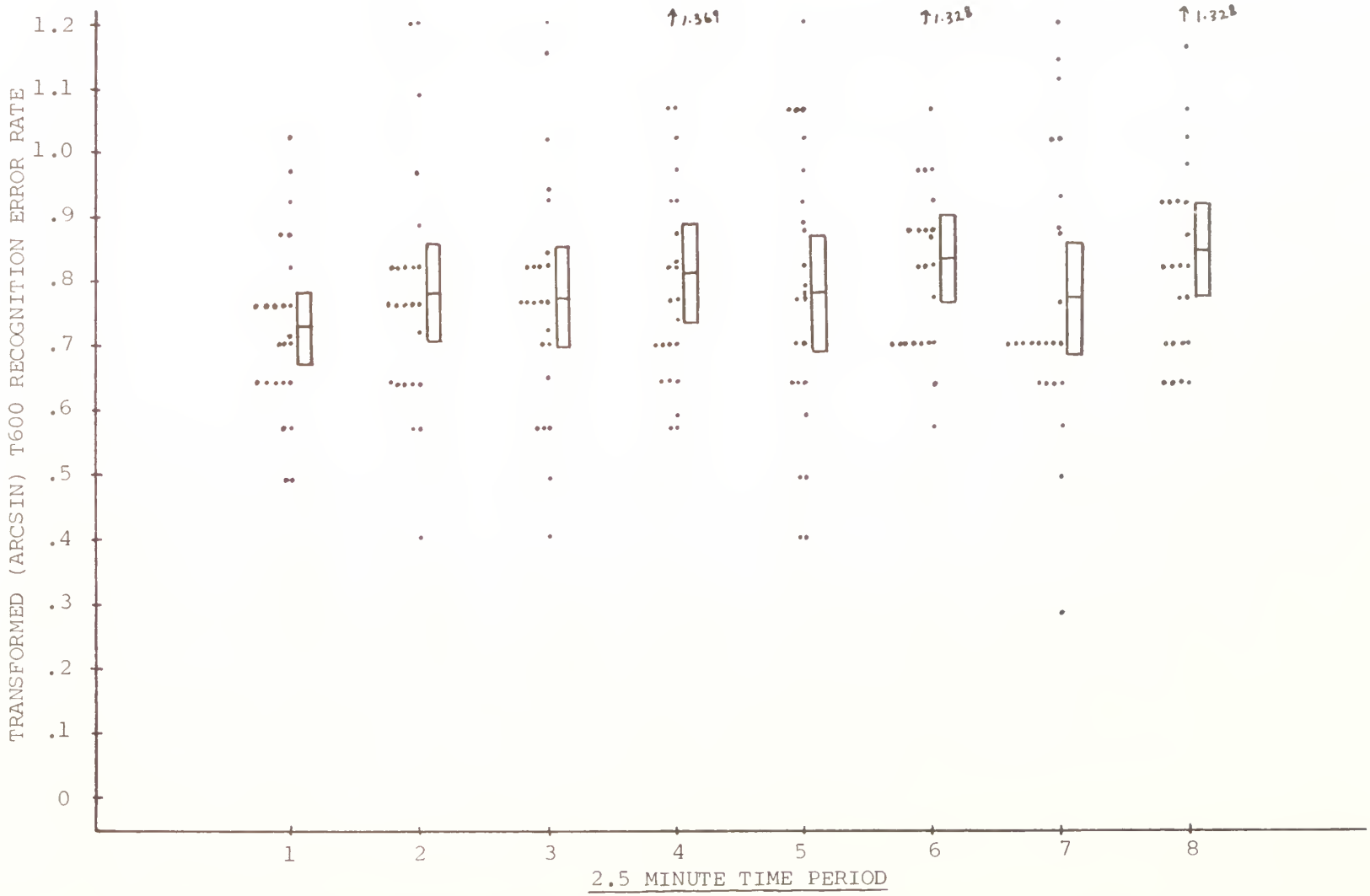
The figure shows a 95% confidence interval for the mean of each 2.5 minute time period. A level of significance, α, of .10 was selected in the experimental design.

FIGURE 6. T600 RECOGNITION ERROR RATE OBSERVATIONS: CONDITION RD1



The figure shows a 95% confidence interval for the mean of each 2.5 minute time period. A level of significance, α, of .10 was selected in the experimental design.

FIGURE 7. TRANSFORMED (ARCSIN) T600 RECOGNITION ERROR RATE OBSERVATIONS: CONDITION NRT



The figure shows a 95% confidence interval for the mean of each 2.5 minute time period. A level of significance, α, of .10 was selected in the experimental design.

FIGURE 8. TRANSFORMED (ARCSIN) T600 RECOGNITION ERROR RATE OBSERVATIONS: CONDITION RD1

TABLE II

ANALYSIS OF VARIANCE FOR T600 RECOGNITION ERROR RATE

Source	df	MS	F
Subjects	23	.218	12.11*
C (Experimental condition)	1	.300	16.67*
T (2.5 minute time period - task duration)	7	.051	2.83**
C x T	7	.008	
Error	345	.018	

* $p < .0005$

** $p < .01$

T_j = contribution of 2.5 minute time period j,

$j = 1,2, \ldots, 8$

e_{ijk} = random error

Subject effects were considered to be random; all others were considered to be fixed.

The analysis showed the experimental condition effect to be significant ($F=16.67$, $df=1/345$, $p < .0005$). The analysis also showed the time period (task duration) effect to be significant ($F=2.83$, $df=7/345$, $p < .01$). A parametric Range Test was performed and concluded that recognition error rates were the same for time periods 2,3,4,5 and 7 and for time periods 6 and 8 and that error rate for time period 1 was less than those for all other time periods ($\alpha = .10$). The interaction between experimental condition and time period was not significant ($F < 1$).

Appendices F, G, H and I present separate confusion matrices for the first, third and eighth 2.5 minute time periods and for all eight 2.5 minute time periods combined respectively. A matrix element a_{ij} of these matrices indicates the proportion of the time that the T600 "thought" that a subject said word j when the subject actually said word i. Figure 9 shows recognition error rate versus 2.5 minute time period for each experimental condition for each vocabulary word type. Figure 10 is a simplified version of figure 9 showing recognition error rate versus 2.5 minute time period for each experimental condition.

The fact that instances arose where subjects either did not speak or spoke a word not in the vocabulary when prompted with a vocabulary word was taken into account in the recognition error rate analysis as it was in the first experiment.

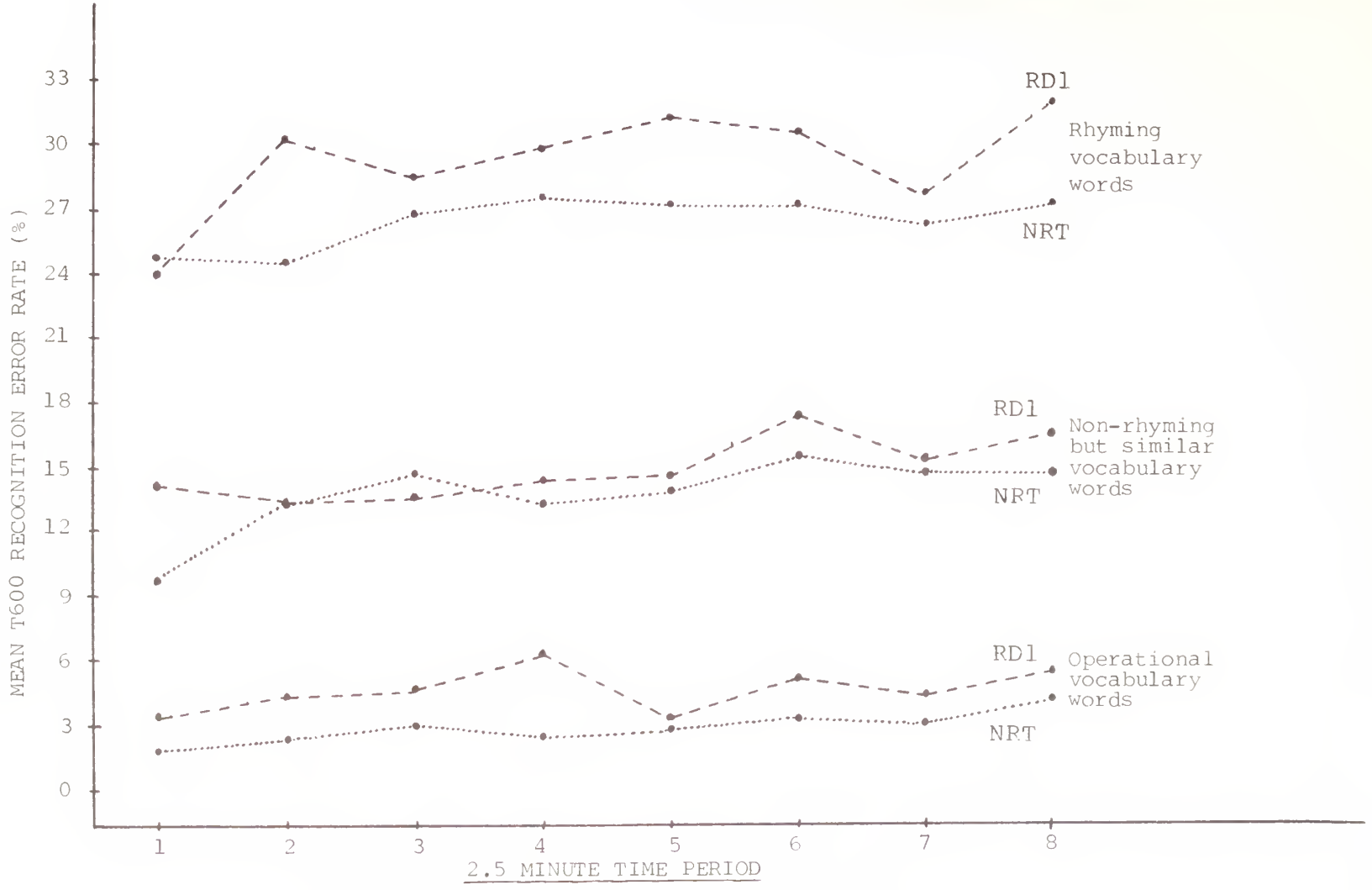


FIGURE 9. MEAN T600 RECOGNITION ERROR RATES

(in recognition errors per 100 spoken utterances)

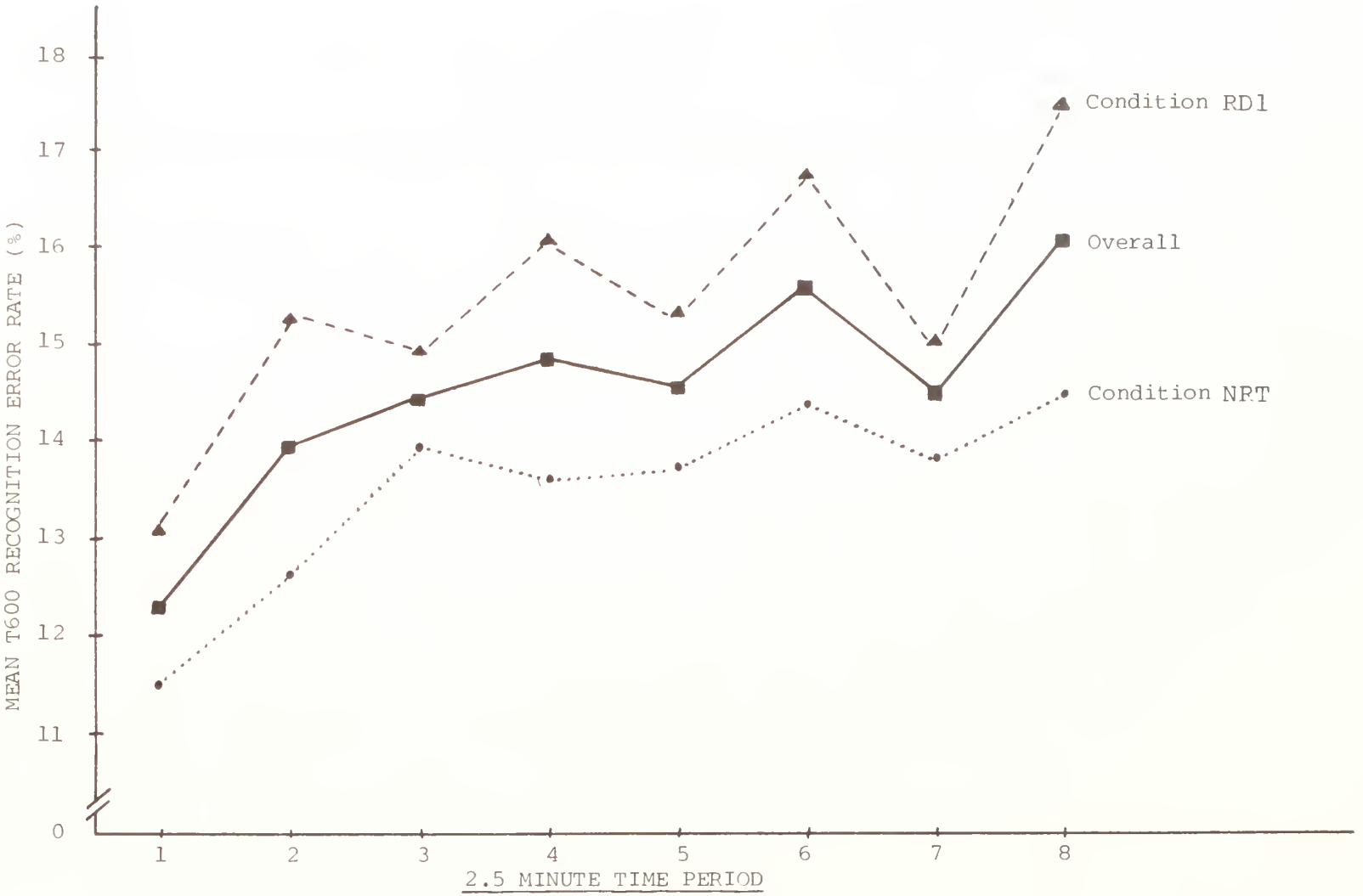


FIGURE 10. MEAN T600 RECOGNITION ERROR RATES
(in recognition errors per 100 spoken utterances)

2. Results for Subject Performance

Appendix J shows total subject verbal errors for each subject for each 2.5 minute time period for each experimental condition. Mean subject verbal error rates for each 2.5 minute time period, experimental condition and vocabulary word type, expressed in subject verbal errors per 100 words presented to the subject for repetition (i.e. each word of the 50 word vocabulary twice), are shown in Table III.

Tests based on the Poisson distribution were performed on the subject verbal error rate data. It was concluded that the 2.5 minute time period effect was not significant ($p > .10$, $\alpha = .10$) and that the experimental condition effect was significant ($p < .0005$, two-tailed test, $\alpha = .10$).

Subject RATER scores for each 2.5 minute time period are shown in Appendix K. A non-parametric Friedman two-way analysis of variance was performed on the RATER scores and it was concluded that the scores were the same for all eight 2.5 minute time periods ($\chi_r^2 = 4.44$, $df = 7$, $p > .7$, $\alpha = .10$).

The results of the subjective fatigue enquiry are shown in Appendix L; numerical scores were obtained as in the first experiment. A non-parametric Wilcoxon matched-pairs signed-ranks test was performed and concluded that subjective fatigue was the same for both experimental conditions ($p > .3$, $\alpha = .10$).

3. General Results

The following were investigated graphically:

a. T600 recognition error rate versus subject verbal error rate; and,

b. RATER scores versus subject verbal error rates.

No relationships were apparent. Spearman rank correlation coefficients between RATER scores and T600 recognition error rates were calculated for each subject; none were found to be significant. (The values calculated ranged from -.446 to .625; r_s(critical) = $\pm$.643, two-tailed test, $\alpha = .10$.)

TABLE III

MEAN SUBJECT VERBAL ERROR RATES*

BY EXPERIMENTAL CONDITION

NRT	.21%
RDI	.74%

BY 2.5 MINUTE TIME PERIOD

1	.38%
2	.71%
3	.38%
4	.25%
5	.63%
6	.42%
7	.38%
8	.67%

BY VOCABULARY WORD TYPE

Rhyming	.36%
Non-rhyming but similar	.56%
Operational	.51%

OVERALL .47%

* Expressed in subject verbal errors per 100 vocabulary words presented to S via the headset. A subject verbal error was defined in this research to be a failure of the subject to repeat correctly the presented vocabulary word. This failure could be either a failure to respond (omission) or responding with a non-vocabulary word or the wrong vocabulary word (commission).

J. DISCUSSION

The results of this experiment support those of Armstrong (1980) and Armstrong and Poock (1981). Subject verbal error rate was higher in experimental condition RD1 than in condition NRT, as expected; corresponding results were obtained in the two experiments cited above. Subject verbal error rate was the same for all eight 2.5 minute time periods; again, corresponding results were obtained in the two experiments cited above. Mean subject verbal error rates in conditions NRT and RD1 (.21% and .74% respectively) appeared less than those observed in the corresponding conditions of Armstrong and Poock (1981), NRT and RD1 (.42% and 1.04% respectively). This was totally unexpected; in fact, mean subject verbal error rates in the conditions of this experiment were expected to be greater than those observed in the corresponding conditions of the first experiment because task duration in this experiment was greater than in the first experiment (20 minutes as opposed to five minutes). Why the observed result occurred is not known.

The subjective fatigue enquiry failed to disclose a significant difference between the two experimental conditions. The result is probably due in part to some subjects scoring the subjective fatigue checklist lower after condition NRT than after condition RD1 because condition NRT bored (under-aroused) them to a great extent whereas condition RD1 aroused them; i.e. (some) subjects and/or the subjective fatigue checklist did not distinguish between fatigue and arousal. (Arousal is discussed later.)

RATER scores did not differ by 2.5 minute time period. This is probably the result of the superposition of learning effects, operating to increase performance with increased task duration, and fatigue effects, operating to decrease performance with increased task duration.

T600 recognition error rate was 15% greater during condition RD1 than during condition NRT. A test based on the Poisson distribution and using just operational vocabulary words also concluded that recognition error rate was greater in condition RD1 than in condition NRT ($p < .0005$). These results substantiate those of the two experiments cited previously regarding increased recognition error rate resulting from increased concurrent tasking (with respect to that experienced during training of the recognizer) and support the generalization of this result to real world vocabularies made in the discussion section of Armstrong (1980).

T600 recognition error rate also differed for some 2.5 minute time periods. This result was not found when using just the operational vocabulary words and a test based on the Poisson distribution ($p > .2$, $\alpha = .10$). Similar results were found in Armstrong and Poock (1981).

Therefore, in this study, mental loading even affected the operational words slightly but the length of time on the task did not.

The results of the range test on recognition error rate differences by 2.5 minute time period and figure 10 suggest that recognition error rate increased quickly during the first five minutes (approximately) of both experimental conditions; further increases occurred at a slower rate, as expected. (Error rate in the eight 2.5 minute period was approximately 30% greater than in the first period.) To substantiate these subjective conclusions, curvilinear regressions (Hicks, 1973) were performed on the recognition error rate versus 2.5 minute time period data for condition NRT, condition RD1 and both conditions combined. In all three cases, the linear terms were significant ($p < .01$). None of the higher order terms were significant (all p's $> \alpha = .10$) but most of the quadratic and cubic terms were noteworthy ($.1 < p < .2$). Inspection of these terms supported the conclusions

reached earlier but also hinted that recognition error rate was starting to increase relatively quickly in the last 2.5 minute time periods. Figure 10 also suggests this. Such an increase was totally unexpected and why it would occur is unknown.

It is again emphasized, as it was by Armstrong (1980) and Armstrong and Poock (1981), that the recognition error rates obtained with the T600 in this experiment are at least ten times what has commonly been found. These higher recognition error rates were deliberately sought by the experimenter (as discussed earlier) and are primarily due to the vocabulary selected.

The results of this experiment are discussed in more detail and interpreted in terms of existing Human Factors Engineering models in the next sections.

K. SUMMARY OF THE THREE RELATED EXPERIMENTS ON MENTAL AND MOTOR LOADING AND TASK DURATION

The following sections will attempt to tie together some of the ideas and results of three experiments which are all related. Those are the studies by Armstrong (1980), Armstrong and Poock (1981) and the current study.

For the reader not familiar with these, the first by Armstrong (1980) examined voice recognition performance while subjects concurrently performed a motor loading tracking task. Three conditions were examined: 1) No tracking task (NTT) 2) Circular tracking task (CTT) and 3) Square tracking task (STT).

Armstrong and Poock (1981) examined the effect of mental loading on voice recognition performance over two - 2 1/2 minute trials. Four levels of mental loading were examined using the RATER device described in this paper to

impose mental loading. The four levels were 1) No rater task (NRT) 2) RATER delay zero (RDØ) 3) RATER delay one (RD1) and 4) RATER delay two (RD2). Based on the results of that study which showed a decrement in performance from the first 2.5 minute trial to the second 2.5 minute trial, the current study was undertaken to examine a longer task duration.

These studies were designed specifically to investigate the effects of task-induced stresses on performance of a voice recognition system comprised of a human operator and a discrete utterance voice recognition system (a Threshold Technology Model T600). Stress was induced through increased concurrent operator mental or motor tasking (with respect to that experienced during training of the recognition device) and task duration. Both 1) increased tasking and 2) task duration were found to affect system performance.

The research was performed in a laboratory setting and employed laboratory tasking devices and a special vocabulary. Consequently, generalizing the results to real world operations should be done cautiously. This research does not predict precisely what will happen in real world settings; however, it does identify two factors which are likely to affect real world operations and gives an indication of the magnitude of the effects of these factors.

The following sections summarize the main results, discuss the results in terms of existing Human Factors Engineering models, and identify further research needed. The following are the main results of the three experiments.

1. T600 recognition error rates (RER) were greater with increased concurrent operator mental or motor tasking (with respect to that experienced during training of the T600). Error rates of conditions RD0, RD1 and

RD2 were 23% greater than the rate of condition NRT (Armstrong and Poock, 1981); those of conditions CTT and STT were 39% greater than that of condition NTT (Armstrong, 1980); and that of condition RD1 was 15% greater than that of condition NRT in the current study. Error rates were the same in conditions RD0, RD1 and RD2 and in conditions CTT and STT. In other words, any amount of mental loading seemed to degrade performance but there was no difference in performance between the various amounts of mental loading themselves. Likewise, any amount of physical motor tracking loading also degraded performance but there was no difference between the tracking tasks themselves.

2. T600 recognition error rates increased with an apparently decreasing rate as task duration increased. This occurred both with and without concurrent mental loading. Similar tendencies were suggested in the experiment with and without concurrent motor loading (Armstrong, 1980) but the differences observed were not statistically significant.
3. Subject verbal error rates were greater with increased concurrent operator mental or motor tasking.
4. Task duration did not affect subject verbal error rates in any of the three experiments.
5. There appeared to be no relationship between subjects' verbal error rates and T600 recognition error rates or performance on the RATER or tracker. Correlation analysis of subjects' recognition error rates relative to performance on the RATER and tracker showed only one significant correlation, between recognition error rate and tracking score in condition STT (square-like path); there was a significant tendency for subjects to do either well or poorly on both the square tracking and voice input tasks but not poorly on one and well on the other.

L. DISCUSSION OF SOME MODELS DESCRIBING THE MAIN RESULTS

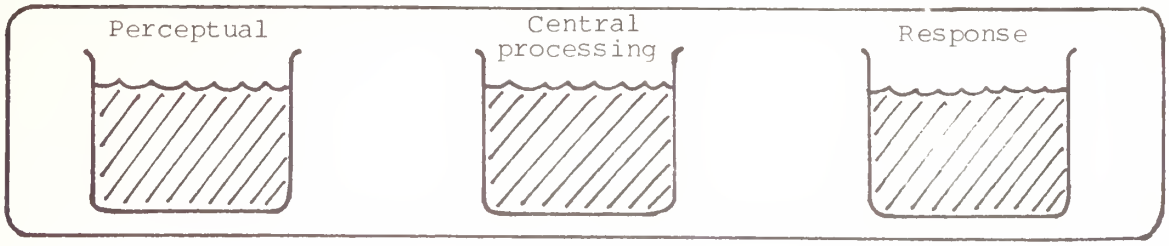
Consideration of the results in terms of several existing Human Factors Engineering models provides some insight into some of the phenomena which were operative. As expected, none of these models is capable of clearly explaining all of the results; most are very specific and hence have limited applicability.

1. Discussion of the T600 Recognition Error Rate (RER) versus Task Loading Results

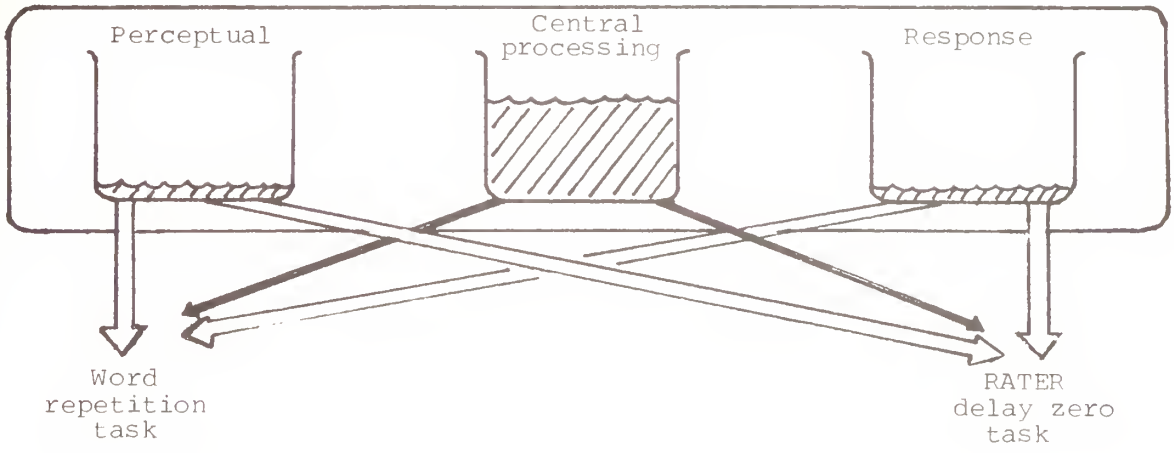
Wickens and Kessel (1979) proposed a model which clearly explains the T600 recognition error rate versus operator task loading results. They consider a human's performance resources to be partitioned into separate structure-specific reservoirs as opposed to one undifferentiated pool of capacity. Various levels of partitioning can be modelled. It suffices here to consider a simple partitioning into the following three distinct reservoirs (Figure 11 - Case A). One is associated with perceptual encoding, i.e. making a categorical classification concerning the nature of a visual or auditory input; another is associated with central processing, an amalgamation of processes involving operations such as rehearsal in short-term memory, risk evaluation and decision making; the third is involved with the execution of behavioral responses. (This description was taken from Wickens and Kessel, 1979.)

Consider the word repetition task in terms of this model. It required perceptual and response resources but very little central processing. Now consider the RATER delay zero task; it also required perceptual and response resources and little central processing. Thus, when the two tasks were performed simultaneously, some competition for perceptual and response resources occurred (Figure 11 - Case B) and consequently performance degraded

CASE A: INITIAL LEVELS (NO TASKING)



CASE B: WORD REPETITION AND RATER DELAY ZERO TASKS



CASE C: WORD REPETITION AND RATER DELAY ONE (OR TWO) TASKS

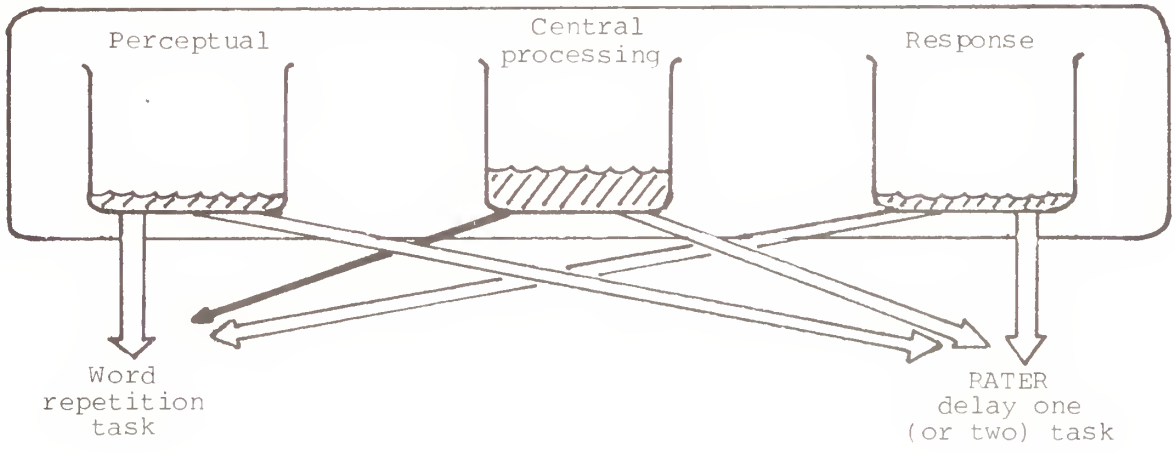


FIGURE 11. HUMAN RESOURCE CAPACITY MODEL
(adapted from Wickens and Kessel, 1979)

(Arrow widths indicate relative resource demands.)

with respect to that observed when the word repetition task was performed separately (i.e. RER(RDO) > RER(NRT)). Now consider the delay one and delay two RATER tasks. They required the same perceptual and response resources as the delay zero task. Although they both required more central processing resources than the delay zero task, these resources were available (Figure 11 - Case C). Consequently, the delay one and delay two tasks did not significantly increase competition for resources beyond that of delay zero and no further performance degradation occurred (i.e. RER(RDO) = RER(RD1) = RER(RD2)).

The circular tracking task also required perceptual and response resources and, because the task was relatively easy and had been practiced, required little central processing. Thus, when it was performed simultaneously with the word repetition task, some competition for perceptual and response resources occurred and recognition error rate increased (i.e. RER(CTT) > RER(NTT)). The square tracking task, however, was more difficult and had not been practiced and consequently required more central processing resources as subjects attempted to learn it; its perceptual and response requirements were similar to those of the circular tracking task. As the additional central processing resources required were available, the competition for resources was approximately the same as with circular tracking and the resulting performance degradation was also the same (i.e. RER(CTT) = RER(STT)).

This interpretation of recognition error rate (RER) versus operator task loading is consistent with the reasoning utilizing the concept of task-induced stress.

The competition for resources brought about by the requirement to perform two tasks simultaneously undoubtedly induced psychological stress; associated changes in subjects' speech then resulted in increased recognition error rate.

2. Discussion of the T600 Recognition Error Rate versus Task Duration Results

The recognition error rate increases observed with increasing task duration were already discussed in this paper in terms of vigilance, learning and fatigue. These increases can also be explained in terms of speech changes due to changes in the level of psychological stress; here the stress level changes are due to fatigue or boredom.

3. Discussion of the Subject Verbal Error Rate versus Task Loading and Task Duration Results

The fact that subject verbal error rates were greater with increased concurrent operator tasking is similar to the recognition error rate versus operator tasking result and is similarly explained by the Wickens and Kessel model. The fact that subject verbal error rate did not differ with task duration was unexpected; it was expected that increased task duration would increase subject verbal error rates as it did T600 recognition error rates. A possible explanation for the observed result is that random variations in the relatively small subject verbal error observations effectively masked any differences which might have existed. This may also be the reason that T600 recognition error rate was found not to depend on task duration when only operational vocabulary words were considered.

4. General Discussion of the Results

The special vocabulary of these experiments was employed specifically as a precaution against the kind of difficulty just discussed (Refer to the Voice Recognition System and Choice of Vocabulary in section D). This mechanism was successful in that it permitted identification of task duration as a factor likely to affect recognition error rate in real world

operations; had just operational type vocabulary words been included in the vocabulary, this potentially important factor might not yet have been recognized.

Arousal theory (as described, for example, in Kling and Riggs, 1971) can also be used to explain most of the results obtained. Arousal is a theoretical construct; its level can be viewed as the theoretical net result of central nervous system activity. Thus, its alleged level at any time is inferred from objective data rather than observed directly. Such inference of the level of arousal can be very subjective and of questionable accuracy.

Behavioral efficiency is hypothesized to have an inverted-U shaped relationship with the behavioral continuum of sleep (under-arousal), wakefulness, and strong excitement (over-arousal); i.e. behavioral efficiency is relatively good for intermediate segments of the behavioral continuum but decreases as distance from the intermediate segment increases in either direction (Figure 12). Most of the results obtained can be explained in terms of this model by simply judiciously hypothesizing the various levels of arousal. For example, consider recognition error rate increases with increased operator loading. If the operator was nearly optimally aroused without concurrent tasking (i.e. near the center of the arousal continuum so that performance was nearly optimal), then increased tasking would have increased his arousal level so that he became over-aroused and his performance degraded.

5. Discussion of the Performance Correlation Results

The arousal model can be employed to explain both the existence of the surprising significant correlation observed between recognition error rate and square-like tracking score (condition STT) and lack of similar

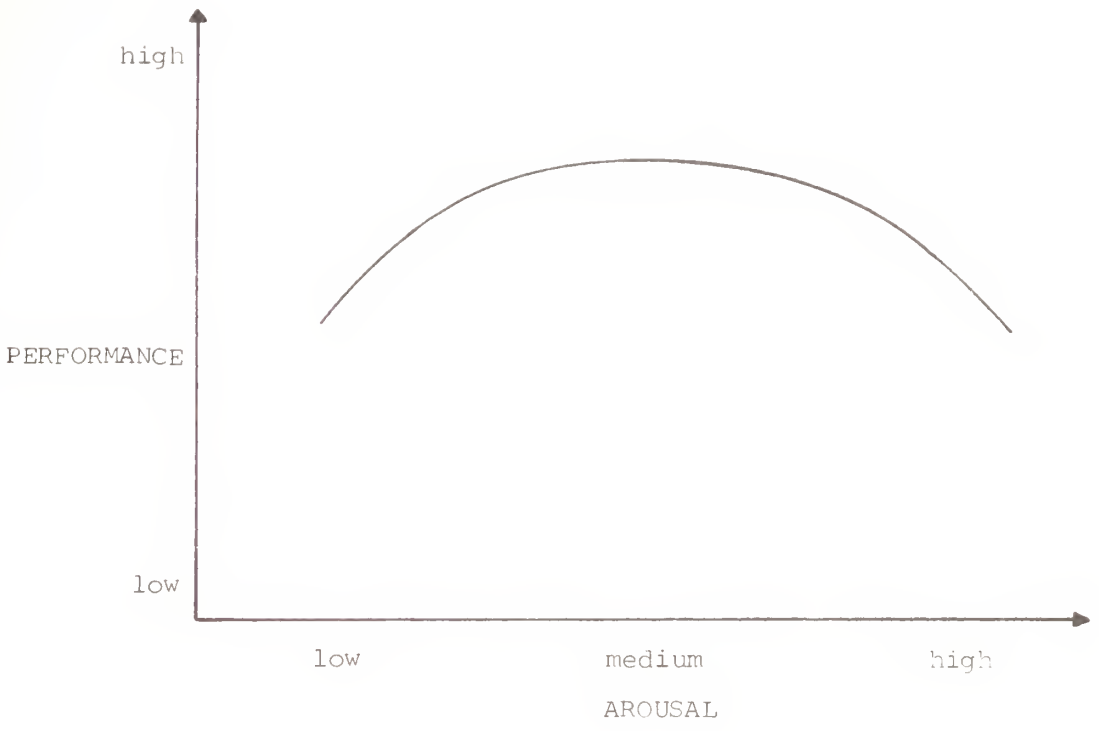


FIGURE 12. PERFORMANCE VERSUS AROUSAL

correlations between recognition error rate and circular tracking score or RATER scores. The value of this is questionable since such explanation requires numerous subjective judgements about relative arousal levels. No really satisfactory explanation of the lone significant correlation observed has been found.

M. FURTHER RESEARCH

As expected, the three studies referred to suggest a myriad of unanswered questions and potentially fruitful follow-up research. Several of these suggested areas of follow-up research are listed below.

1. Investigate the possibility that, when prompted orally, subjects may mimic to some extent the prompting speech. If such a tendency exists, it could conceivably be confounded with the task loading and time on task effects investigated.
2. Investigate performance of a recognition system with concurrent tasking which induces higher levels of stress or interferes more with the voice task than the current research. The Wickens and Kessel model discussed earlier should be useful in designing this research.
3. Investigate the observed task duration effect over longer time periods than used in this research.
4. Investigate the interaction between the stressing task and performance of the voice recognition system at a microscopic level as compared to the macroscopic view adopted in this research, i.e. investigate the extent to which instantaneous performance on the stressing task affects recognition performance at the same instant. It seems intuitive that if the stressing task is causing a subject great difficulty at any particular instant then performance

on the recognition task would degrade compared to when the stressing task was causing the subject less difficulty.

5. Investigate training the recognizer under operator task-loading conditions similar to those that will be experienced during operation. This would parallel Drennan's (1980) and Elster's (1981) research into training and operating a recognition system under various ambient noise levels.

6. Investigate the existence of a functional relationship between performance of the voice recognition system and operator task loading at any time that can be used to predict the task loading given the level of performance of the recognition system. This potential use of voice recognition system performance to monitor operator workload and/or stress was suggested by Feuge and Geer (1978).

N. CONCLUDING REMARKS

This research contributes to filling in some of the known gaps in voice recognition technology. In addition, it used a new technique in voice recognition research, use of a special vocabulary to facilitate detection of factors influencing system performance. Two factors which may adversely affect performance of real world voice recognition systems were identified: 1) task duration, and 2) increased concurrent operator mental or motor workload (with respect to that experienced during training of the voice recognition system). The performance degradations associated with these factors were not known before. How much these factors will affect any particular real world operation will depend on the specifics of that operation. The results using the special vocabulary also indicated the unique performance that real world operators may encounter when using very similar phrases, such as "above glide slope" and "below glide slope" or "VHF" and "UHF".

REFERENCES

1. Armstrong, J. W. The effects of concurrent Motor Tasking on performance of a voice recognition system. Master's Thesis, Naval Postgraduate School, 1980.

2. Armstrong, J. and Poock, G. K. Effect of Operator Mental Loading on Voice Recognition System Performance. Naval Postgraduate School Tech Rept # NPS55-81-016, 1981.

3. Brady, J. F., Perceptual-motor Performance Degradation as a Function of Passively Induced Vestibular Coriolis Stimuli, General Dynamics, Convair Division, Report GDC-ERR-1287, 1968.

4. Davies, D. R. and Tune, G. S., Human Vigilance Performance, American Elsevier Publishing Company, Inc., 1969.

5. Elster, R. S. The Effects of Certain Background Noises on the Performance of a Voice Recognition System. Naval Postgraduate School Technical Report NPS54-80-010, September 1980.

6. Hicks, C. R., Fundamental Concepts in the Design of Experiments, Second Edition, Holt, Rinehart and Winston, 1973.

7. Kennedy, R. S. The relationship between habituation to vestibular stimulation and vigilance: Individual Differences and Subsidiary Problems. Naval Aerospace Medical Research Lab, Report NAMRL-Mono-20, 1972.

8. Kling, J. W. and Riggs, L. A., Woodworth and Schlosberg's Experimental Psychology, Third Edition, Holt, Rinehart and Winston, Inc., 1971.

9. Kryter, K. D., in Van Cott, H. P. and Kinkade, R. G. (editors), Human Engineering Guide to Equipment Design, Revised Edition, 1972.

10. Long, G. M. and Fishburne, R. P., Performance Norms and Research Applications of the Response Analysis Tester (RATER), Naval Aerospace Medical Research Laboratory, Aerospace Psychology Technical Memorandum 73-1, 7 September 1973.

11. Martin, T. B. and Grunza, E. F., Voice Control Demonstration System, Air Force Avionics Laboratory, Report TR-74-174, (AD-B004 928L) March 1974.

12. Newsom, B. D., Brady, J. F. and O'Laughlin, T. W., Study of Performance in a Revolving Space Station Simulator as a Function of Head Rotation about Y and Z Cranial Axes, General Dynamics, Convair Division, Report GDC DBD 65-043-12, 21 November 1966.

13. Pearson, R. G. and Byars, G. E. Jr., The Development and Validation of a Checklist for Measuring Subjective Fatigue, Air University School of Aviation Medicine, USAF, Report 56-115, December 1956.

14. Poock, G. K., Experiments with Voice Input for Command and Control, Naval Postgraduate School, Technical Report NPS-55-80-016, April 1980.

15. Scheffé, H., The Analysis of Variance, John Wiley and Sons, Inc., 1959.

16. Scott, P. B., Alpha-numeric Extraction Technique, Rome Air Development Center, Report TR-75-287, (AD-B008 303L) November 1975.

17. Scott, P. B., Word Recognition, Rome Air Development Center, Report TR-78-209, (AD-A061 545) September 1978.

APPENDIX A

VOCABULARY LISTING (BY WORD TYPE)

RHYMING

gale	tale	gold	cold
game	came	bark	park
tip	dip	big	pig
beat	peat	ten	den

NON-RHYMING BUT SIMILAR

sap	sat	peas	peace
race	raze	save	safe
lake	late	kit	kid
mad	mat		

OPERATIONAL

list	course	attack	refuel
time	plot	bingo	cancel
speed	air	report	proceed
dive	fire	distance	label
drop	launch	copy	station

A vocabulary listing in the order in which the words were trained is attached to the written instructions initially given to subjects and shown in Appendix C.

APPENDIX B

SUBJECT DATA SHEET

Subject number: _____ Name: _____ Age: _____

Time/date: _____ Service: _____

Rank: _____ MOS (in words): _____

Do you object to being taperecorded during the experiment? If you do, stop filling out this form and advise the experimenter now; otherwise, continue.

How many hours experience have you had on voice recognition equipment in the last six months?

_____ hours (approximately)

How many hours experience have you had on reaction measurement devices in the past year?

_____ hours (approximately)

Do you have a speech or hearing impediment? Yes No
(circle one)

Do you want a post participation briefing on your performance and on the hypotheses being tested by the experimenter? Note that if you request such a briefing, you must agree not to discuss this with anyone other than the experimenter so that no subject will learn what results are expected prior to his participation in the experiment; such prior knowledge could invalidate the results of the experiment.

Yes No
(circle one)

After you have completed participation in the experiment you will be asked to write below any comments which you think may be useful to the experimenter. If you have any questions now, please ask the experimenter. Otherwise, give him this form now and start reading the pages titled "INTRODUCTORY REMARKS/ RECOGNIZER VOCABULARY TRAINING".

POST EXPERIMENT COMMENTS

(continue on reverse side if this space is insufficient)

THANK YOU FOR YOUR PARTICIPATION

APPENDIX C

WRITTEN INSTRUCTIONS

INTRODUCTORY REMARKS / RECOGNIZER VOCABULARY TRAINING

INTRODUCTORY REMARKS

This experiment involves analysis of a combined human operator / voice recognition equipment system under various conditions of operator mental loading. The actual experiment will be carried out in a sound-proof booth and subject - experimenter communication during the actual experiment will be via the booth intercom system; however, you may remove the headset assembly during break periods and leave the booth.

CAUTION: The mounting of the voice recognizer microphone on the headset assembly is very delicate, easily damaged, and difficult to repair. Please be careful while handling this assembly.

Please carry out the experiment exactly as directed and do not discuss your performance with anyone other than the experimenter as inappropriate subject prior knowledge could invalidate the results.

VOICE RECOGNIZER VOCABULARY TRAINING

The 50 word vocabulary being used with the voice recognizer in this experiment is attached to these instructions. You will be required to repeat each word of this vocabulary ten times to train the recognizer to recognize your particular vocalizations of each word. To facilitate recognition by the voice recognizer, you should include in the ten repetitions

as many as possible of the different ways you might say the word in normal speech; for example, use different intonations and emphasis, and small variations in volume.

In order to keep track of the number of times you say each word, and to reduce breath noise, it is best to speak the 10 repetitions in several groups. For example, if the word is zero, it is better to group them as:

	000-000-0000
or	000-000-000-0
rather than as	0000000000
or	0-0-0-0-0-0-0-0-0-0

Please observe the following guidelines while inputting voice data to the recognizer both during training and later during the actual experiment.

a. Speak each word crisply and quickly but do not over-pronounce; for example, words ending in "t" - delete final "t" if more natural.

b. Be sure to leave a distinct pause (specifically, at least one-tenth of a second of silence) between each word so that the recognizer can distinguish the end of one word from the beginning of the next. Similarly, do not leave a period of silence within a word or the recognizer will mistake it for two separate words.

c. Avoid breathing into the microphone at the end of words as this will generate false inputs to the recognizer.

d. Microphone location is very important and should be kept constant throughout the experiment; i.e., adjust it if it gets out of place. The experimenter will initially demonstrate correct microphone placement.

From this point on instructions will be given to you verbally by the experimenter. Please advise him if you have any questions now.

VOCABULARY LISTING (IN TRAINING ORDER)

0. attack
1. list
2. gale
3. tale
4. bingo
5. sap
6. sat
7. time
8. gold
9. cold
10. cancel
11. peas
12. peace
13. speed
14. game
15. came
16. distance
17. race
18. raze
19. copy
20. bark
21. park
22. launch
23. save
24. safe
25. refuel
26. tip
27. dip
28. drop
29. lake
30. late
31. course
32. big
33. pig
34. report
35. kit
36. kid
37. plot
38. beat
39. peat
40. proceed
41. mad
42. mat
43. fire
44. ten
45. den
46. label
47. air
48. station
49. dive

APPENDIX D

SUBJECTIVE FATIGUE CHECKLIST

Subject number _____ Experimental condition _____

FEELING TONE CHECK LIST

No.	Better than	Same as	Worse than	Statement
1.	( )	( )	( )	slightly tired
2.	( )	( )	( )	like I'm bursting with energy
3.	( )	( )	( )	extremely tired
4.	( )	( )	( )	quite fresh
5.	( )	( )	( )	slightly pooped
6.	( )	( )	( )	extremely peppy
7.	( )	( )	( )	somewhat fresh
8.	( )	( )	( )	petered out
9.	( )	( )	( )	very refreshed
10.	( )	( )	( )	ready to drop
11.	( )	( )	( )	fairly well pooped
12.	( )	( )	( )	very lively
13.	( )	( )	( )	very tired

Have you checked each statement?

INSTRUCTIONS FOR COMPLETING FEELING TONE CHECKLIST

People feel different at various times for various reasons. Some arise after a night's rest feeling "quite rested" while others may feel "a little tired". A hard day's work or a vigorous workout at the gym may make you feel "fairly well pooped"; yet, a shower, a cup of coffee, or merely a few minutes relaxing in a comfortable chair may make you feel "very refreshed".

I would like to find out how you feel right now. On the accompanying sheet, you will see 13 statements which describe different degrees of freshness or peppiness and tiredness. For each statement you will have to determine in your own mind whether you feel at this instant (1) "Better than", (2) the "Same as", or (3) "Worse than" the feeling described by that statement. Having done this you will then place an "X" in the appropriate box.

Consider the following example:

No.	Better than	Same as	Worse than	Statement
0.	( )	( )	( )	somewhat tired

If right now you felt "somewhat tired" you would place an "X" in the box marked "Same as". If, however, you felt fresh or full of pep you would check the box marked "Better than" because you would be feeling better than "somewhat tired". On the other hand, if you felt exhausted you would place an "X" in the box marked "Worse than".

Take each statement in order; do not skip around from one to another. Read each statement carefully so that you understand what it means. It may help you to understand some statements if you mentally insert the words "I feel" or "I am" before the statement.

This is not a test. You have all the time you need.

APPENDIX E

T600 RECOGNITION ERRORS*

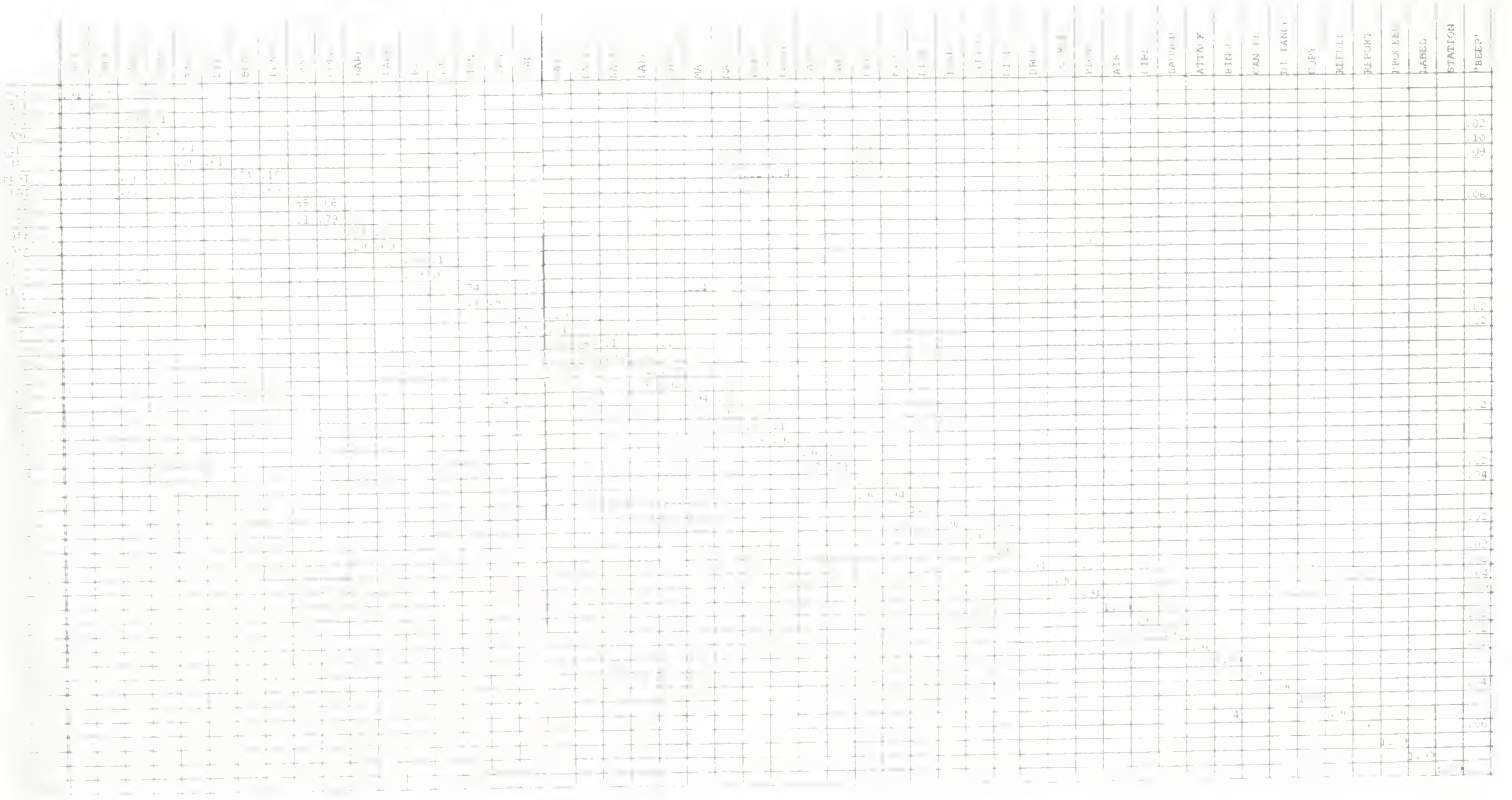
An entry x/y indicates that a total of x recognition errors occured and that y of these were errors of rejection (of vocabulary words).

SUBJECT NUMBER	EXPERIMENTAL CONDITION NRT — 2.5 MINUTE TIME PERIOD								EXPERIMENTAL CONDITION RD1 — 2.5 MINUTE TIME PERIOD							
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	7/0	3/0	3/3	7/2	7/4	10/4	11/5	9/1	7/1	4/1	7/0	5/1	5/0	4/0	6/0	5/0
2	5/1	3/1	9/1	7/1	6/0	4/0	7/1	8/0	3/0	5/0	8/1	9/1	13/0	10/1	9/0	12/2
3	5/1	7/1	7/1	6/2	9/1	8/1	5/0	8/0	7/1	8/2	7/2	7/1	7/0	6/0	7/2	9/2
4	3/0	7/1	6/1	4/0	10/1	7/0	6/1	5/2	7/2	13/4	8/5	11/3	7/2	11/4	12/2	11/2
5	1/0	8/0	7/0	5/1	6/0	7/0	5/0	5/0	6/0	8/5	7/2	6/2	13/3	8/2	5/1	8/1
6	2/1	5/0	5/0	6/0	6/0	3/0	5/0	6/0	4/0	4/0	4/0	8/1	3/0	7/1	5/1	6/0
7	3/0	7/0	11/1	5/0	7/0	4/0	5/0	1/0	3/0	5/0	5/1	6/0	7/0	6/0	4/0	6/0
8	4/0	5/2	6/1	5/1	4/2	6/2	5/2	7/1	5/1	7/1	3/1	5/1	2/0	5/1	6/1	8/2
9	1/0	6/0	5/0	5/0	3/0	6/1	3/0	4/0	10/1	11/1	10/1	8/0	9/3	6/1	10/0	10/0
10	6/0	3/1	3/0	5/2	4/1	7/2	10/4	7/1	5/0	8/0	4/0	5/0	6/1	6/0	6/1	7/1
11	7/0	12/2	9/1	14/2	14/3	13/5	8/2	11/1	11/6	16/5	16/6	12/5	13/2	13/2	9/2	15/4
12	5/1	5/0	8/1	6/1	5/0	3/0	8/1	8/0	5/0	5/0	4/0	4/0	4/0	9/0	6/0	5/0
13	5/0	3/0	4/0	6/0	7/0	7/0	6/0	7/1	7/0	8/0	8/0	6/2	6/0	6/0	5/0	5/0
14	13/5	7/3	12/4	6/1	5/1	12/2	8/3	11/5	5/1	7/0	7/0	6/1	8/0	8/0	5/0	8/0
15	8/0	9/0	7/0	9/0	8/0	8/0	6/0	11/1	8/0	9/0	12/2	10/0	12/1	11/0	12/1	10/2
16	11/0	8/1	10/2	14/1	10/2	10/0	10/2	10/1	9/1	16/0	15/1	20/2	16/0	19/0	16/1	13/1
17	6/0	5/0	3/0	3/0	4/0	6/2	7/0	6/0	7/0	2/0	2/0	4/0	2/0	6/0	3/0	6/0
18	5/0	3/0	7/0	3/0	5/0	6/0	9/0	5/0	5/0	7/1	6/0	6/0	3/0	8/0	6/1	5/0
19	6/0	5/0	9/0	9/2	6/0	4/0	12/1	10/1	4/0	6/0	8/0	13/1	11/1	9/0	6/0	10/1
20	6/0	11/1	9/0	14/2	11/1	13/2	5/1	10/4	12/2	5/1	10/0	10/1	10/1	11/1	14/0	19/1
21	10/2	8/2	8/2	7/2	6/2	6/1	4/1	7/1	6/1	8/2	7/0	7/2	5/0	8/0	1/0	7/1
22	6/0	9/1	8/0	5/0	6/1	8/0	8/2	4/1	6/0	7/0	8/1	6/0	5/0	9/1	6/0	6/0
23	8/2	10/4	5/1	7/4	9/3	9/3	8/1	7/1	6/1	8/2	6/1	13/4	7/2	6/1	6/1	8/2
24	5/0	2/0	6/0	2/0	7/1	6/0	5/1	8/0	9/0	5/0	6/4	8/2	9/0	9/1	14/3	10/0

* A T600 recognition error was operationally defined in this research to be a failure of the T600 to recognize correctly any vocabulary word which S spoke and includes both incorrect recognition and rejection of vocabulary words; T600 recognition errors do not include those cases where S spoke a word not in the vocabulary (or coughed, sighed, etc.) and the T600 generated a recognition.

APPENDIX F

CONFUSION MATRIX FOR 2.5 MINUTE TIME PERIOD NUMBER 1

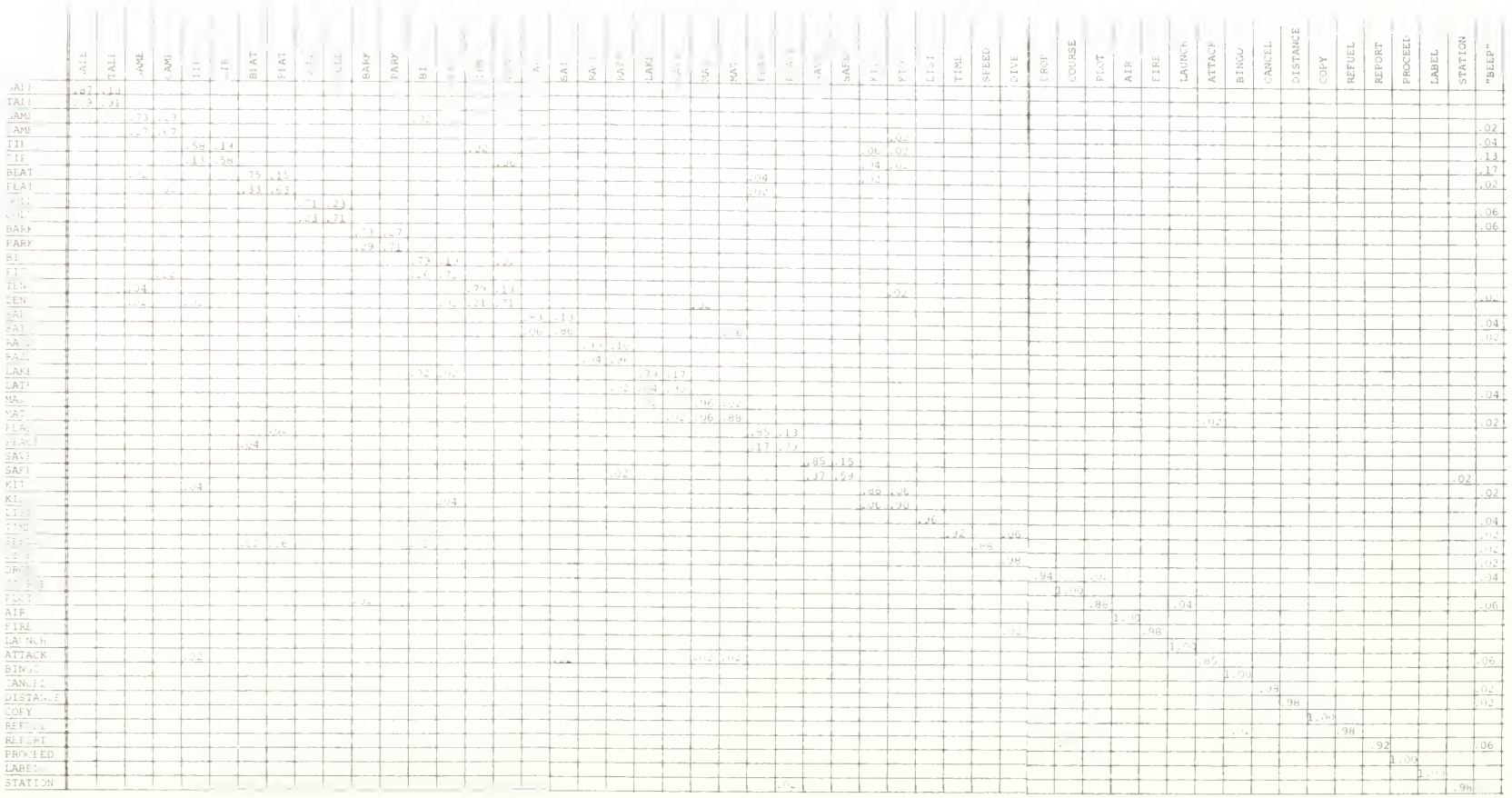


APPENDIX G

CONFUSION MATRIX FOR 2.5 MINUTE TIME PERIOD NUMBER 3

The matrix element i,j indicates the proportion of the time that the [illegible] thought" that a subject said word j when the subject actually said word i. The last column, annotated "beep", indicates the proportion of the time that the T600 rejected vocabulary word i and "beeped".

Several rows do not sum to exactly [illegible] this is because of truncation in the elements of the row as well, amounting to two decimal places.

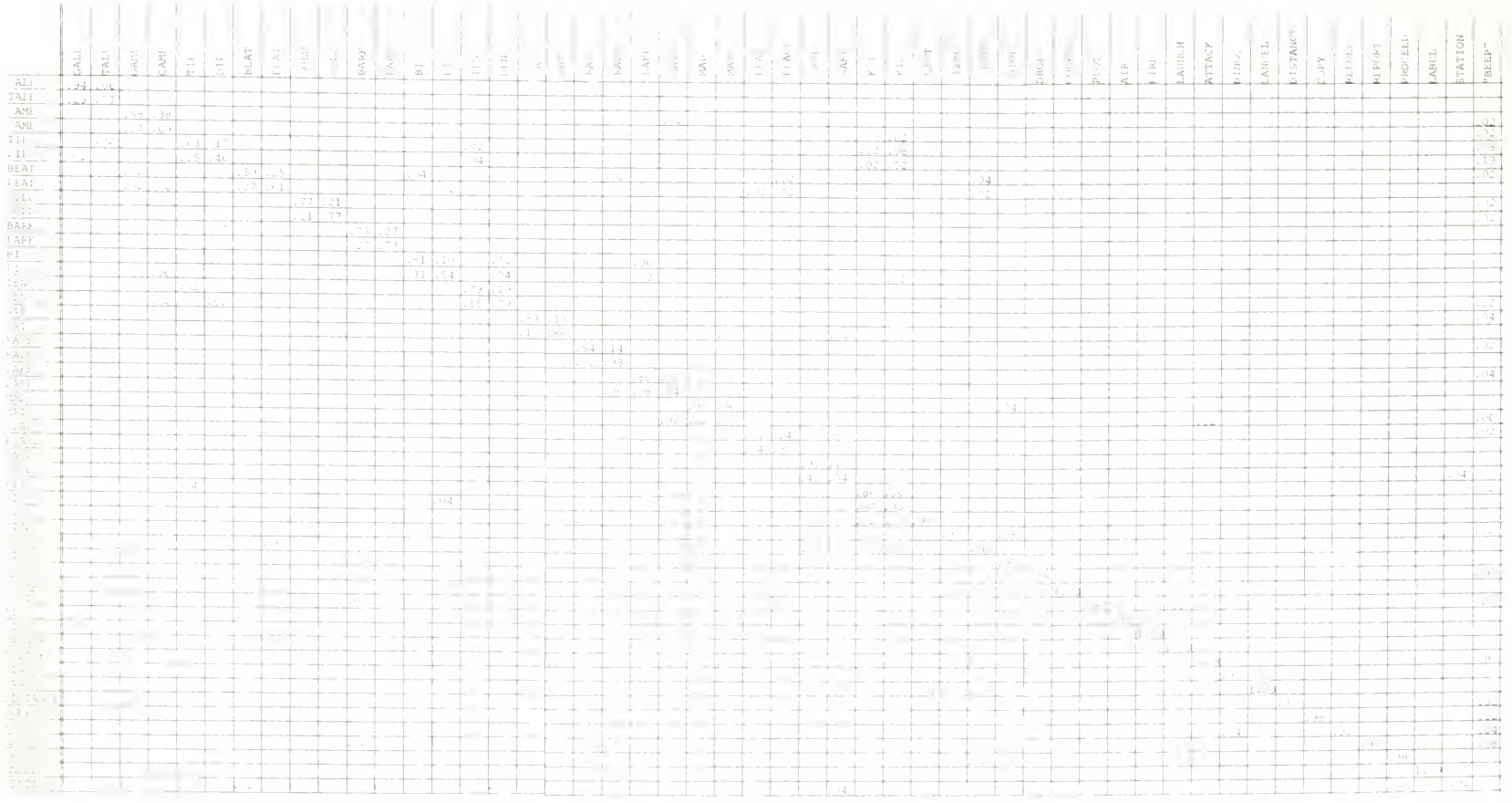


APPENDIX H

CONFUSION MATRIX FOR 2.5 MINUTE TIME PERIOD NUMBER 8

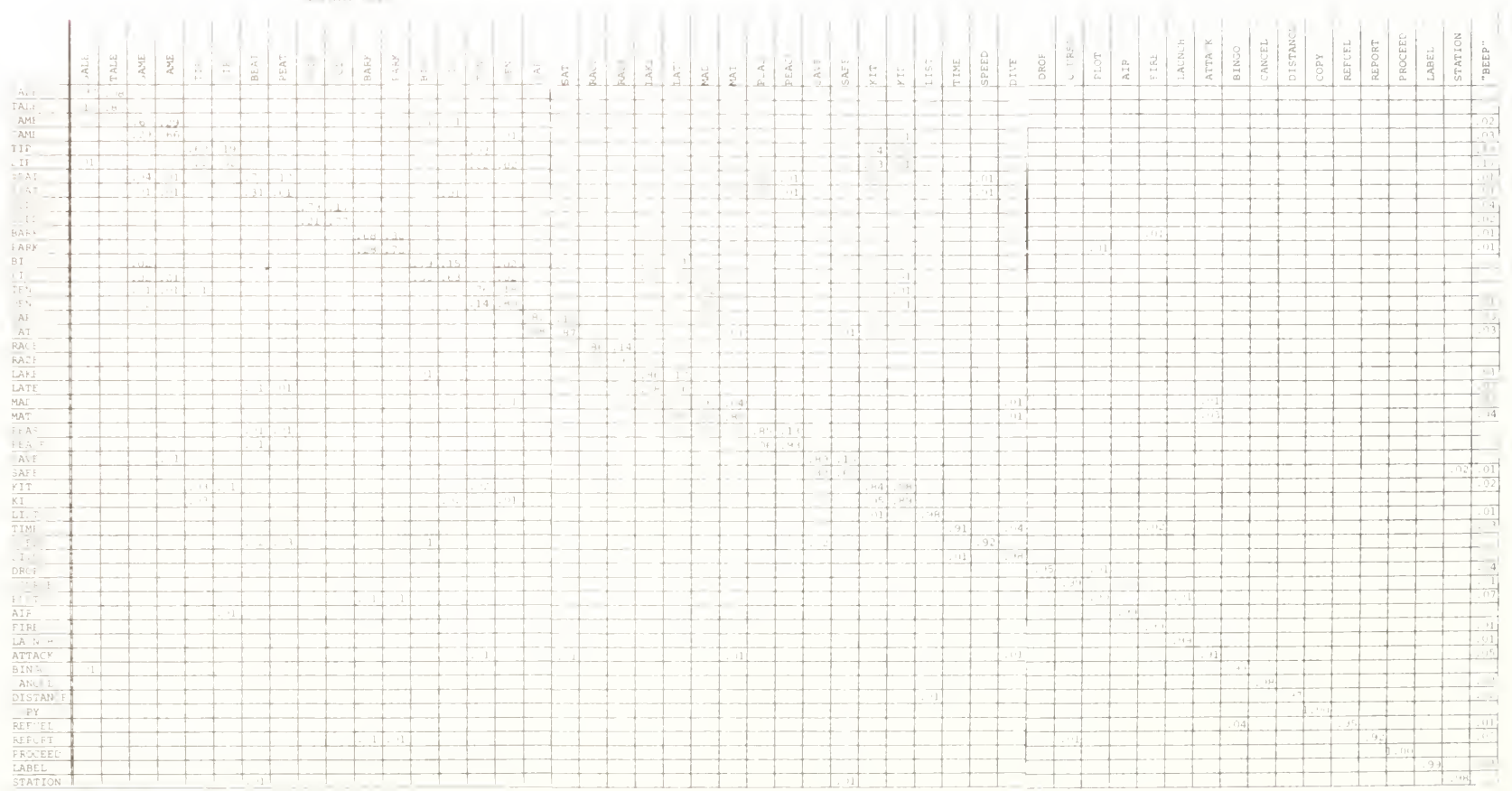
The matrix element [illegible] indicates the proportion of the time that the [illegible] said word [illegible] when the [illegible] actually said word [illegible]. The last column, entitled "[illegible]", indicates the proportion of the time that the TRW rejected the vocabulary word [illegible] and "beeped".

Several [illegible] of the matrix [illegible] rounding to two decimals.



APPENDIX I

CONFUSION MATRIX FOR ALL EIGHT 2.5 MINUTE TIME PERIODS COMBINED



SUBJECT VERBAL ERRORS*

An entry x/y indicates that a total of x subject verbal errors occured and that y of these were errors of not speaking any word or speaking a non-vocabulary word (when prompted with a vocabulary word).

SUBJECT NUMBER	EXPERIMENTAL CONDITION 2.5 MINUTE TIME PERIOD								EXPERIMENTAL CONDITION 2.5 MINUTE TIME PERIOD							
	1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
2	1/1	0/0	0/0	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0	0/0
3	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	1/1	0/0	0/0	3/3	1/1	1/1	0/0
4	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	1/0	0/0	1/1	0/0	0/0	1/1
5	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	0/0	1/1	0/0	1/0
6	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0
7	0/0	0/0	0/0	0/0	0/0	1/0	0/0	1/0	0/0	0/0	1/1	0/0	1/0	0/0	0/0	1/0
8	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	0/0	0/0	0/0	1/1
9	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	1/1	0/0	2/2	0/0
10	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	1/1
11	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	2/1
12	0/0	0/0	0/0	0/0	1/1	0/0	0/0	2/1	0/0	0/0	0/0	0/0	1/1	3/2	1/0	1/0
13	0/0	1/1	1/1	0/0	1/1	0/0	0/0	1/1	0/0	1/1	0/0	0/0	0/0	0/0	0/0	0/0
14	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0
15	0/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	0/0	0/0	0/0	0/0
16	0/0	0/0	0/0	0/0	0/0	0/0	1/0	1/0	1/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0
17	0/0	1/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	0/0	0/0	0/0	0/0
18	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	1/1	2/2	0/0	1/0
19	0/0	1/1	0/0	0/0	1/1	0/0	0/0	0/0	0/0	1/1	0/0	0/0	0/0	0/0	0/0	0/0
20	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/0	0/0	0/0	0/0	1/1	0/0	0/0
21	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0
22	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	3/3	1/1	2/2	4/4	1/1	0/0	0/0	0/0
23	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	0/0
24	1/1	0/0	0/0	0/0	0/0	0/0	0/0	0/0	1/1	2/2	3/3	1/1	1/1	0/0	1/1	1/1

* A subject verbal error was defined in this research to be a failure of a subject to repeat correctly the presented vocabulary word. This failure could be either a failure to respond (omission) or responding with a non-vocabulary word or the wrong vocabulary word (commission).

APPENDIX K

RATER SCORES

SUBJECT NUMBER	2.5 MINUTE TIME PERIOD							
	1	2	3	4	5	6	7	8
1	98	100	100	94	98	99	100	93
2	95	90	91	80	82	91	87	91
3	98	88	98	87	92	96	95	97
4	90	87	76	96	98	88	88	77
5	86	83	85	87	91	87	78	89
6	97	82	88	84	75	92	96	92
7	72	91	83	86	89	91	95	100
8	89	84	79	94	91	90	92	88
9	94	88	93	91	98	94	80	98
10	98	100	97	100	96	100	98	100
11	82	61	84	73	84	79	85	76
12	91	88	97	89	85	80	91	88
13	72	77	81	74	82	80	66	60
14	100	90	94	18	91	91	87	98
15	99	98	97	98	99	96	99	100
16	77	89	92	92	88	92	98	71
17	88	94	94	98	93	92	96	96
18	75	82	78	73	79	73	85	72
19	96	98	94	86	99	93	91	89
20	100	99	90	100	100	84	100	100
21	72	69	62	72	71	77	73	76
22	73	77	92	55	91	93	90	91
23	92	98	98	96	90	95	97	94
24	82	77	75	96	97	88	89	88
Mean	88.2	87.1	88.3	84.1*	90.0	89.2	89.8	88.5

* The relatively low mean PATER score in time period number four is due to the two outliers underlined. Post-experiment discussion with the two subjects concerned revealed that the subjects probably inadvertently and temporarily switched delay modes.

A perfect score for any 2.5 minutes time period was 100.

APPENDIX L

SUBJECTIVE FATIGUE SCORES*

SUBJECT NUMBER	EXPERIMENTAL CONDITION ST	DT
1	21	13
2	10	11
3	15	14
4	12	12
5	11	12
6	14	14
7	12	14
8	14	13
9	13	14
10	13	13
11	10	11
12	16	12
13	12	12
14	8	7
15	15	14
16	12	12
17	12	12
18	12	10
19	16	10
20	13	14
21	12	16
22	13	16
23	12	12
24	13	13

* Higher scores are associated with lower subjective fatigue and vice versa.

Scores were calculated as in the first experiment

APPENDIX M

VOICE RECOGNITION STUDIES AT NPS

This project is one of several voice recognition research projects conducted for/by Professor G. K. Poock at NPS over the last several years. The complete list, in addition to this report, includes:

Armstrong, J. W., The Effects Of Concurrent Motor Tasking On Performance Of A Voice Recognition System, Masters Thesis, Naval Postgraduate School, Monterey, 1980.

Armstrong, J. W. and Poock, G. K. Effect of Operator Mental Loading on Voice Recognition System Performance. Naval Postgraduate School Technical Report NPS55-81-016, 1981.

Batchellor, M. P., Investigation Of Parameters Affecting Voice Recognition Systems In C^3 Systems, Masters Thesis, Naval Postgraduate School, Monterey, 1981.

Bragaw, P. H., Investigation Of Voice Input For Constructing Joint Chiefs Of Staff Emergency Action Messages, Masters Thesis, Naval Postgraduate School, Monterey, 1981.

Elster, R. S. The Effects Of Certain Background Noises On the Performance Of a Voice Recognition System, Naval Postgraduate School Report NPS54-80-010, September 1980.

Jay, G. T., An Experiment In Voice Data Entry for Imagery Intelligence Reporting, Masters Thesis, Naval Postgraduate School, Monterey, 1981.

McSorley, W. J. Using Voice Recognition Equipment To Run The Warfare Environmental Simulator (WES), Masters Thesis, Naval Postgraduate School, Monterey, 1981.

Neil, D. E. and Andreason, T. Examination Of Voice Recognition System To Function In a Bilingual Mode, Naval Postgraduate School Report NPS55-81-003, February 1981.

Poock, G. K. Experiments With Voice Input For Command And Control: Using Voice Input To Operate A Distributed Computer Network, Naval Postgraduate School Report NPS55-80-016, April 1980.

Poock, G. K. A Longitudinal Study of Computer Voice Recognition Performance and Vocabulary Size, Naval Postgraduate School Technical Report NPS55-81-013, June 1981.

Poock, G. K. To Train Randomly Or All At Once ... That Is The Question. Proceedings of Voice Data Entry Systems Applications Conference, Sponsored by Lockheed Missiles and Space Co., Sunnyvale, California, Oct. 7-8, 1981.

Taggart, J. L. and Wolfe, C. D., Speech Recognition As An Input Medium For Preflight In The P3C Aircraft, Masters Thesis, Naval Postgraduate School, Monterey, 1981.

DISTRIBUTION LIST

	No. of Copies
DEFENSE TECHNICAL INFORMATION CENTER CAMERON STATION ALEXANDRIA, VA 22314	2
LIBRARY, CODE 0142 NAVAL POSTGRADUATE SCHOOL MONTEREY, CA 93940	2
DEAN OF RESEARCH CODE 012 NAVAL POSTGRADUATE SCHOOL MONTEREY, CA 93940	1
LIBRARY, CODE 55 NAVAL POSTGRADUATE SCHOOL MONTEREY, CA 93940	1
PROFESSOR GARY POOCK, CODE 55PK OPERATIONS RESEARCH DEPARTMENT NAVAL POSTGRADUATE SCHOOL MONTEREY, CA 93940	234